GENE-ENVIRONMENT INTERACTIONS IN GENETIC EPIDEMIOLOGY

A Dissertation

by

CHRISTINE M. SPINKA

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

December 2004

Major Subject: Statistics

GENE-ENVIRONMENT INTERACTIONS IN GENETIC EPIDEMIOLOGY

A Dissertation

by

CHRISTINE M. SPINKA

Submitted to Texas A&M University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Approved as to style and content by:

Raymond J. Carroll (Chair of Committee)	
Ruzong Fan (Member)	Joanne Lupton (Member)
Bani K. Mallick (Member)	Michael Longnecker (Head of Department)

December 2004

Major Subject: Statistics

ABSTRACT

Gene-Environment Interactions in Genetic Epidemiology. (December 2004)
Christine M. Spinka, B.S. Vanderbilt University; M.S., Texas A&M University
Chair of Advisory Committee: Dr. Raymond J. Carroll

Gene-environment interactions are an area of increasing interest in complex human diseases. The first step in any study of the interactions between genes and the environment involves identifying genes which influence the trait of interest. In this dissertation, a new method for using the information in complex pedigrees to perform a joint linkage disequilibrium and linkage mapping of quantitative trait loci is developed. Subsequently, methods are needed to determine the interaction, if any, between these genes and environmental risk factors. Many of these factors, such as weight or age, are continuous and little is known about their distributions. Thus, we introduce a new method for estimating the gene-environment interaction parameters in a logistic regression for the case-control study design. In doing so, we make the assumption that in the underlying population, the distributions of the genetic factors and the environmental covariates are independent. Additionally, we treat the environmental parameters nonparametricly, utilizing the profile likelihood. Furthermore, the methodology we develop is also general enough to be used on many different types of genetic information, including haplotypes, and can accommodate missing genotype data. The method is also extended to allow analysis in the presence of population stratification or genotype misclassification. We show that the standard errors of parameter estimates using our method are smaller than those found using complete data

only. These methods are illustrated using simulations and are applied to a real data set exploring the interaction between genotype and environment in disease risk.

To my loving family

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Raymond Carroll, for countless hours of his time and assistance. Under his direction, I have learned much about undertaking research and writing papers. Furthermore, I would like to thank Dr. Carroll for providing me many intellectually enriching opportunities. For example, these have included participation in the bioinformatics training program as well as collaborations with scientists at the National Cancer Institute and in the Department of Nutrition at Texas A&M University. Additionally, I would like to thank Dr. Nilanjan Chatterjee for opening my eyes to an interesting research question and providing me access to a motivating data set. Finally, I would like to thank my family and friends for all of their love and support. Without you, this work could never have become a reality.

TABLE OF CONTENTS

Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1 Background

Many important human diseases arise from the joint actions of a variety of genetic and environmental factors. Diseases such as cancers, osteoporosis, and diabetes are just a few examples of human diseases which arise through this mechanism. Gaining a better understanding of the underlying causes, both genetic and environmental, will help scientists and physicians to better treat disease.

Human subjects provide a unique set of challenges to scientists seeking to study disease formation. The ethics and logistics of human research require specialized study designs. The need for specialized study designs arises in part from the fact that humans may only be studied with their consent; neither their behaviors nor matings can be controlled. Thus, methodologies must be developed that can account for the particular data types that are often be collected in human studies.

Within the framework of human studies, there are two main goals which must be addressed. First, genes or genetic locations which are associated with risk of disease must be identified. This is often a difficult problem, as many of the diseases of interest are effected by tens or hundreds of disease locations (loci) within the genome. Second, once important genes are identified, it is necessary to understand both how the gene influences the formation of diseases and how it interacts with environmental factors.

The format and style follow that of *Journal of the American Statistical Association*.

1.2 Mapping Quantitative Trait Loci Using Complex Pedigrees

Many methods have been proposed to allow the identification of important locations in the genome which are associated with a particular quantitative trait. These locations are called quantitative trait loci or QTLs. Methods to identify QTLs often rely upon a particular data structure, for example nuclear families, or sibling pairs. This dependence upon a particular data structure often makes data collection difficult and results in a large fraction of available data not being used in the analysis. For example, affected sibling pair methods can only use the data from two siblings; a single offspring with one affected parent cannot be used. Thus, methods which can incorporate the information in all types of family structures can prove more powerful and less expensive than their more restrictive counterparts.

Previous work in this area has included the development of variance component methods to identify regions in the genome which are associated with disease risk. Methods which utilize population data are often used to localize QTLs to a particular region or regions of the genome. However, these methods are susceptible to population stratification which can suggest an association where there is none or hide a true region. To combat this problem, many researchers use family based study methodologies.

Recently (Fan and Jung (2003); Fan and Xiong (2002); Fan and Xiong (2003)) developed a method to perform a joint linkage and linkage disequilibrium mapping of quantitative trait loci (QTLs). Their method utilizes population data and data from nuclear families or sibling pairs together to identify potential QTLs in the genome. This method is extended in this dissertation to allow incorporation of family data from any pedigree structure. Two particular extended pedigrees are used to illustrate the method. Simulation results are provided to illustrate the performance of the

method.

1.3 Gene-environment Interactions

Once important locations in the genome have been identified, it is of interest to gain an understanding of how these genes interact with the environment in the formation of disease. Currently there are several methods available for this type of study; many of these methods use data from related individuals, or from prospective studies, which can be difficult or expensive to perform. The method proposed here allows gene-environment interactions to be studied using case-control study data. These studies have a variety of advantages for human studies; they do not require related individuals to study, they can be performed on historical data, and they can be easier to implement and much less expensive than family or prospective studies.

In the context of case-control studies and gene-environment interactions, we address three main types of situations. First, we consider the case where individuals are sampled from a homogeneous population. In this case, it is assumed that a simple case-control study is performed, and that all covariates are measured without error. However, the genotype may be partially or entirely unknown. Therefore, the method of Prentice and Pyke(1979) can not be used in this situation, as it applies to the complete data setting only. On the other hand, Chatterjee and Carroll present a method which assumes marginal independence between genotype and environment. In this dissertation, their method is extended to the present case in which the probability of disease in the population may be either known or unknown. Finally, the results are compared for two cases. This method is particularly applicable to haplotypes and thus simulations for this type of data are also presented.

We also consider the problem of estimating gene-environment interactions in the context of a stratified population. These populations occur when each individual

considered belongs to one of several sub-populations, often racial or cultural groups, which have different joint distributions of the factors of interest. In the presence of population stratification, naive estimates of the genetic, environmental, and interaction effects can show spurious associations. These problems have caused case-control studies to fall out of favor with some scientists. However, when variables are measured that help identify the strata, we propose a method to consistently estimate the gene-environment interactions, as well as genetic and environmental effects. This method is illustrated using a simulation study, which indicates that the method performs well, even in the presence of very distinct sub-populations.

Additionally, we examine the effects of genetic misclassification on procedures of this type and develop methodology to analyze data having this structure. Genetic misclassification is a well known problem in the literature, and arises when attempts to genotype an individual provide incorrect results. For example, Wong et al.(2004) show that in the presence of misclassification, regression estimates of gene-environment interaction effects are biased. We develop a similar method for case-control studies that provides unbiased estimates of model parameters. The results of a simulation study are presented to illustrate the method.

Finally, we use the methods developed in this dissertation to analyze the data from a case-control study of ovarian cancer. The data set contains variables which are believed to influence the development of ovarian cancer in Israeli women. Examples of these factors include the presence or absence of a mutation at a particular location and the use of oral contraceptives. The methods developed here are especially appropriate for this data, as over half of all of the study participants have unknown genotype.

CHAPTER II

JOINT LINKAGE DISEQUILIBRIUM AND LINKAGE MAPPING FOR COMPLEX PEDIGREES

2.1 Introduction

Mapping quantitative trait loci (QTL) for complex diseases may be performed using linkage disequilibrium (LD) regression analysis on population data. However, the presence of population substructures may affect the results and either produce false positives or mask true associations. To combat this problem, variance component models have been proposed to perform joint LD and linkage mapping of QTLs using both population and pedigree data (Abecasis, Cookson, and Cardon (2001); Allison et al. (1998); Almasy et al. (1999); Cardon (2000); Fan and Jung (2003); Fan and Xiong (2003); Fulker et al. (1999); Göring and Terwilliger (2000); Martin et al. (2000); Sham et al. (2000)). Unfortunately, many of the current methods are limited to data sets comprised of small nuclear families. This restriction makes data collection difficult and costly when the trait of interest is rare.

Previous works (Fan and Jung (2003); Fan and Xiong (2002); Fan and Xiong (2003)) have utilized either family or sibling data in combination with population data to perform joint linkage and LD mapping of QTLs. Their work is generalized in this chapter to allow the incorporation of multi-generational pedigrees involving relatives of any type. Intuitively, large pedigrees contain more linkage and LD information than simple nuclear families. Thus, it is important to develop models that include all types of data, including population data, sib-ships, nuclear families and multi-generational pedigrees in a combined analysis.

This chapter is organized as follows. In Section 2.2 we introduce variance compo-

nent models for multi-generational pedigree data. These models include both linkage and LD parameters; their genetic effects are decomposed into an orthogonal summation of additive and dominant components. The method is compared with the "AbAw" approach (Abecasis et al. 2000, 2001; Cardon (2000); Fulker et al. (1999); Sham et al. (2000)). For consistency and ease of comparison, the pedigrees found in Figure 1 of Abecasis, Cookson, and Cardon (2000) are considered as examples. In Section 2.3 we present analytical formulas that approximate the non-centrality parameters of the proposed test statistics; these are compared with Abecasis, Cookson, and Cardon (2000). Lastly, in Section 2.4 we provide some concluding remarks. Proofs are provided in Appendix A.

2.2 Methods

In this section, the method to perform a joint analysis of LD and linkage information is outlined. First, consider a biallelic quantitative trait locus Q which has alleles Q_1 and Q_2 with frequencies q_1 and q_2, respectively. Assume that markers A and B are typed and located in the same chromosomal region as the trait locus Q. Additionally, let markers A and B each have two alleles A and a or B and b with frequencies P_A and P_a or P_B and P_b, respectively, and let the two alleles combine under Hardy-Weinberg equilibrium. Assume that the data set is composed of I independent families with n_i individuals in the i-th family; denote them by $j = 1, 2, \cdots, n_i$, where each individual j has a larger index than all of his ancestors. Denote the quantitative traits of i-th family by a vector $\mathbf{y}_i = (y_{i1}, y_{i2}, \cdots, y_{in_i})^\tau$, their genotypes at marker A by a vector $(A_{i1}, A_{i2}, \cdots, A_{in_i})^\tau$, and their genotypes at marker B by a vector $(B_{i1}, B_{i2}, \cdots, B_{in_i})^\tau$.

Now, define G_{ij} to be the polygenic effect and e_{ij} to be the random error, where

$G_{ij} \sim \mathrm{N}(0, \sigma_G^2)$, $e_{ij} \sim \mathrm{N}(0, \sigma_e^2)$, and the two are independent. Further, define

$$x_{Aij} = \begin{cases} 2P_a & \text{if} A_{ij} = AA \\ P_a - P_A & \text{if} A_{ij} = Aa \\ -2P_A & \text{if} A_{ij} = aa \end{cases}, \qquad z_{Aij} = \begin{cases} -P_a^2 & \text{if} A_{ij} = AA \\ P_a P_A & \text{if} A_{ij} = Aa \\ -P_A^2 & \text{if} A_{ij} = aa \end{cases}$$

Define x_{Bij} and z_{Bij} similarly. Note that the x's and z's are a transformation of the genotypes to isolate the additive and dominant effects, respectively. We can then model the value of the quantitative trait, similar to Fan and Jung (2003), as

$$y_{ij} = \beta + x_{Aij}\alpha_A + x_{Bij}\alpha_B + z_{Aij}\delta_A + z_{Bij}\delta_B + G_{ij} + e_{ij}, \tag{2.1}$$

where $\mu = (\beta, \alpha_A, \alpha_B, \delta_A, \delta_B)^\tau$ is the vector of parameters.

Now, decompose the total variance, σ^2, as $\sigma^2 = \sigma_g^2 + \sigma_G^2 + \sigma_e^2$, where σ_g^2 is the variance explained by the QTL Q, σ_G^2 is the polygenic variance, and σ_e^2 is the error variance. Both σ_g^2 and σ_G^2 can be further decomposed into their additive and dominant components, $\sigma_g^2 = \sigma_{ga}^2 + \sigma_{gd}^2$ and $\sigma_G^2 = \sigma_{Ga}^2 + \sigma_{Gd}^2$. Let π_{jkQ} be the proportion of alleles shared identical by descent (IBD) at QTL Q by the j-th and the k-th individuals, and Δ_{jkQ} be the probability that both alleles at QTL Q are shared IBD by the j-th and the k-th individuals. π_{jkQ} and Δ_{jkQ} are usually estimated by marker information, see for example Amos (1994), Amos and Elston (1989), or Amos et al. (1989). The estimates of π_{jkQ} and Δ_{jkQ} are functions of the recombination fractions (Almasy and Blangero (1998); Fan and Jung (2003); Fulker et al. (1995); Goldgar and Oniki (1992); Pratt et al. (2000)) and thus, linkage information is included in the variance-covariance matrix. Now, define Σ_i to be the variance-covariance matrix for family i, where the

(j,k)th element of Σ_i is

$$\rho_{jk} = \sigma^2 \begin{cases} 1 & j = k \\ (\pi_{jkQ}\sigma^2_{ga} + \Delta_{jkQ}\sigma^2_{gd} + \sigma^2_{Ga}/2 + \sigma^2_{Gd}/4)/\sigma^2 & j < k \cdot \\ (\pi_{kjQ}\sigma^2_{ga} + \Delta_{kjQ}\sigma^2_{gd} + \sigma^2_{Ga}/2 + \sigma^2_{Gd}/4)/\sigma^2 & k < j \end{cases}$$

Using this setup, the design matrix for family i can be written as $X_i = (X_{i1}, ..., X_{in_i})^\tau$ where $X_{ij} = (1, x_{Aij}, x_{Bij}, z_{Aij}, z_{Bij})^\tau$. This allows us to write the log-likelihood for family i as $L_i = -\frac{n_i}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(\mathbf{y}_i - X_i\mu)^\tau\Sigma_i^{-1}(\mathbf{y}_i - X_i\mu)$.

Let μ_{ij} be the effect of genotype $Q_iQ_j, i, j = 1, 2$, $\mu_{12} = \mu_{21}$, where $(\mu_{11}+\mu_{22})/2 = 0$, as in traditional quantitative genetics, see for example Falconer and Mackay (1996). Then, $a = \mu_{11} = -\mu_{22}$, $d = \mu_{12}$, $\alpha_Q = q_1\mu_{11} + (q_2 - q_1)\mu_{12} - q_2\mu_{22} = a + (q_2 - q_1)d$ is the average allele substitution effect, and $\delta_Q = 2\mu_{12} - \mu_{11} - \mu_{22} = 2d$ is the dominant effect. Under this setup, the additive variance $\sigma^2_{ga} = 2q_1q_2\alpha^2_Q$ and the dominant variance $\sigma^2_{gd} = (q_1q_2)^2\delta^2_Q$.

Denote the linkage disequilibrium between to loci, A and B by $D_{AB} = P(AB) - P_AP_B$; then the LD between Q and A is $D_{QA} = P(AQ_1) - q_1P_A$, and the LD between Q and B is $D_{QB} = P(BQ_1) - q_1P_B$. Define the additive and dominant variance-covariance matrices as $V_A = \begin{pmatrix} 2P_aP_A & 2D_{AB} \\ 2D_{AB} & 2P_bP_B \end{pmatrix}$ and $V_D = \begin{pmatrix} P^2_aP^2_A & D^2_{AB} \\ D^2_{AB} & P^2_bP^2_B \end{pmatrix}$. Then, Fan and Xiong (2002) show that the coefficients of regression equation (2.1) are given by

$$\begin{pmatrix} \alpha_A \\ \alpha_B \end{pmatrix} = V_A^{-1}\begin{pmatrix} 2D_{AQ} \\ 2D_{QB} \end{pmatrix}\alpha_Q \text{ and } \begin{pmatrix} \delta_A \\ \delta_B \end{pmatrix} = V_D^{-1}\begin{pmatrix} D^2_{AQ} \\ D^2_{QB} \end{pmatrix}\delta_Q.$$

Thus, linkage effects the variance-covariance matrix, while association effects the mean coefficients. Then, tests of either linkage or association can be performed by

comparing the full model in which all parameters are estimated to a sub-model in which some parameters are set equal to zero.

If both additive and dominant variances σ^2_{ga} and σ^2_{gd} are significantly larger than 0, a null hypothesis $H_{AB,ad} : \alpha_A = \alpha_B = \delta_A = \delta_B = 0$ can be tested. However, if the additive variance, σ^2_{ga}, is significantly larger than 0, but dominant variance, σ^2_{gd}, is not significantly larger than 0, then regression (2.1) can be simplified by excluding the dominant effects from the analysis, i.e., setting $\delta_A = \delta_B = 0$. Then, the null hypothesis $H_{AB,a} : \alpha_A = \alpha_B = 0$ may be tested.

2.3 Non-centrality Parameter Approximations

In order to assess the performance of the suggested likelihood ratio tests, approximate forms for the non-centrality parameters are needed. In Fan and Jung (2003) the values of these parameters for nuclear families and sibling pairs have been determined. However, we show that asymptotic values for the forms of the non-centrality parameters can be found for any pedigree structure, using tools from linear regression in combination with the tables found in Appendix A.

To illustrate this method for extended pedigrees, we consider I families of the form given either in graph A or graph B of Figure 1. Note that these are the same pedigrees as in Abecasis, Cookson, and Cardon (2000), Figure 1. Let N be the total number of individuals, i.e., $N = nI$, where $n = 11$ for graph A and $n = 18$ for graph B. Let $\mathbf{y} = (\mathbf{y}_1^\tau, \cdots, \mathbf{y}_1^\tau)^\tau$; then $\mathbf{y}$ is normal with mean $X\mu$, where $X = (X_1^\tau, \cdots, X_I^\tau)^\tau$, and variance-covariance matrix $\Sigma = diag(\Sigma_1, \cdots, \Sigma_I)$. Assume that $\Sigma_1 = \cdots = \Sigma_I$. Now, let $\hat{\mu}$ be the maximum likelihood estimators of μ then, $\hat{\mu} = \left[X^\tau \hat{\Sigma}^{-1} X\right]^{-1} X^\tau \hat{\Sigma}^{-1} \mathbf{y} = \left[\sum_{i=1}^I X_i^\tau \hat{\Sigma}_i^{-1} X_i\right]^{-1} \sum_{i=1}^I X_i^\tau \hat{\Sigma}_i^{-1} \mathbf{y}_i$. Next, when H is any $q \times 5$ test matrix of rank q, using linear model theory the null hypothesis $H\mu = 0$

is testable, see Graybill (1976). Furthermore, the test statistic, F,

$$F = \frac{(H\hat{\mu})^\tau [H(X^\tau \hat{\Sigma}^{-1} X)^{-1} H^\tau]^{-1} (H\hat{\mu})}{\mathbf{y}^\tau [\hat{\Sigma}^{-1} - \hat{\Sigma}^{-1} X (X^\tau \hat{\Sigma}^{-1} X)^{-1} X^\tau \hat{\Sigma}^{-1}] \mathbf{y}} \frac{N-5}{q},$$

is distributed as a non-central $F(q, N-5, \lambda)$, where the non-centrality parameter has the form

$$\lambda = (H\mu)^\tau \Big[H[X^\tau \Sigma^{-1} X]^{-1} H^\tau \Big]^{-1} H\mu = (H\mu)^\tau \Big[H[\sum_{i=1}^{I} X_i^\tau {\Sigma_i}^{-1} X_i]^{-1} H^\tau \Big]^{-1} (H\mu).$$

Denote $\Sigma_i^{-1} = \frac{1}{\sigma^2}(\gamma_{kl})_{n\times n}$. In Appendix A, we develop an approximation of the non-centrality parameter,

$$\lambda \approx \frac{I}{\sigma^2} (H\mu)^\tau \left[H \begin{pmatrix} 1/\sum_k \sum_l \gamma_{kl} & O & O \\ O & V_A^{-1}/b_1 & O \\ O & O & V_D^{-1}/b_2 \end{pmatrix} H^\tau \right]^{-1} (H\mu),$$

where b_1 and b_2 are calculated according to complex formulas found in the appendix and O denotes the zero matrix.

Consider the null hypothesis $H_{AB,a} : \alpha_A = \alpha_B = 0$; denote the corresponding F-test statistic by $F_{AB,a}$. Then, the non-centrality parameter has the form $\lambda_{AB,a} \approx \frac{b_1 I}{\sigma^2} \sigma_{ga}^2 [P_b P_B D_{AQ}^2 - 2 D_{AQ} D_{AB} D_{QB} + P_a P_A D_{QB}^2] / [q_1 q_2 (P_a P_A P_b P_B - D_{AB}^2)]$. Similarly, consider the null hypothesis $H_{AB,d} : \delta_A = \delta_B = 0$; denote the corresponding test statistic by $F_{AB,d}$. In this case, the non-centrality parameter is $\lambda_{AB,d} \approx \frac{b_2 I}{\sigma^2} \sigma_{gd}^2 [P_b^2 P_B^2 D_{AQ}^4 - 2 D_{AQ}^2 D_{AB}^2 D_{QB}^2 + P_a^2 P_A^2 D_{QB}^4] / [q_1^2 q_2^2 (P_a^2 P_A^2 P_b^2 P_B^2 - D_{AB}^4)]$. Finally, define the null hypothesis $H_{AB,ad} : \alpha_A = \alpha_B = \delta_A = \delta_B = 0$; denote the corresponding test statistic by $F_{AB,ad}$. Then, the associated non-centrality parameter is $\lambda_{AB,ad} = \lambda_{AB,a} + \lambda_{AB,d}$.

Assume that only one marker, A, is used in the analysis. Then, the test statistic for the null hypothesis $H\mu = 0$ is distributed as non-central $F(q, N-3)$, where $\mu = (\beta, \alpha_A, \delta_A)^\tau$. The non-centrality parameter for the null hypothesis $H_{A,ad} : \alpha_A =$

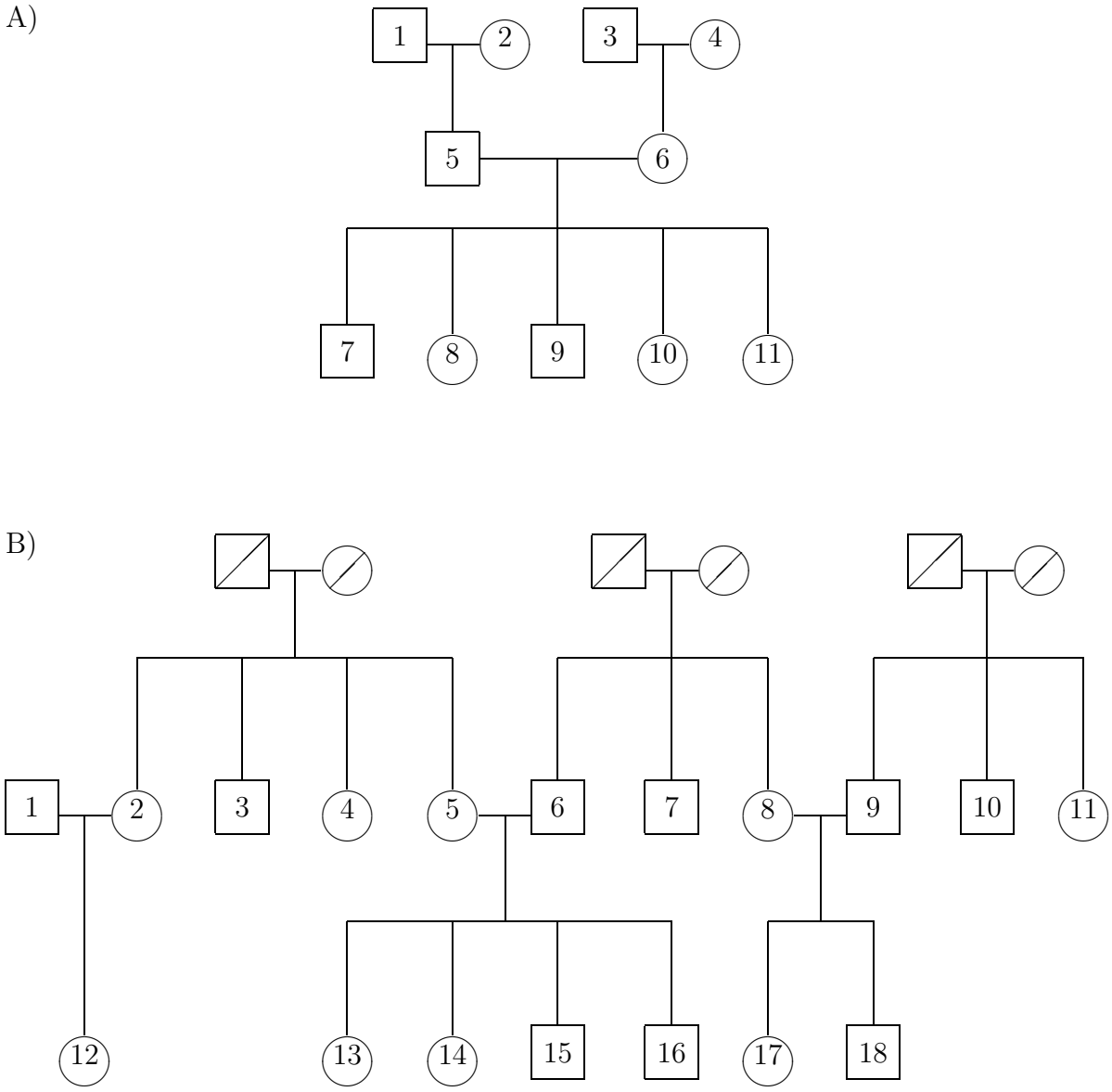

Figure 1. Pedigrees used in power calculations and comparison, which are taken from Figure 1 of Abecasis, Cookson, and Cardon (2000). The number in the box or circle is individual ID.

$\delta_A = 0$ is then $\lambda_{A,ad} \approx \frac{b_1 I}{\sigma^2}\sigma^2_{ga} D^2_{AQ}/(P_A P_a q_1 q_2) + \frac{b_2 I}{\sigma^2}\sigma^2_{gd} D^4_{AQ}/(P^2_A P^2_a q^2_1 q^2_2)$. Similarly, the non-centrality parameter for the null hypothesis $H_{A,a} : \alpha_A = 0$ has the form $\lambda_{A,a} \approx \frac{b_1 I}{\sigma^2}\sigma^2_{ga} D^2_{AQ}/(P_A P_a q_1 q_2)$. The non-centrality parameter for the null hypothesis $H_{A,d} : \delta_A = 0$ is $\lambda_{A,d} \approx \frac{b_2 I}{\sigma^2}\sigma^2_{gd} D^4_{AQ}/(P^2_A P^2_a q^2_1 q^2_2)$.

2.3.1 Power Comparisons with the "AbAw" Approach

Table 1 shows the power of 50 duplicates of pedigree A) and pedigree B), for varying levels of LD between the trait locus and marker A at 0.01 significant level. The parameters selected are the same as those of Abecasis, Cookson, and Cardon (2000). Additionally, the power of χ^2_{qtl} is taken from Table 2 of Abecasis, Cookson, and Cardon (2000). Let $F_(A, a)$ denote the test statistic for the null hypothesis $H_{A,d} : \delta_A = 0$. The power of $F_{A,a}$ is calculated using the non-centrality parameter approximation $\lambda_{A,a}$. As can be seen in Table 1, $F_{A,a}$ has higher power than χ^2_{qtl}. However, Abecasis, Cookson, and Cardon (2000) show that χ^2_{qtl} is more powerful than other methods, and thus our proposed method is advantageous over the "AbAw" approach.

Table 1. Power (%) of 50 pedigrees (Figure 1) for varying levels of LD between trait locus and marker A at 0.01 significant level. The parameters are given by $\sigma^2_{ga} = 0.1, \sigma^2_{gd} = 0, \sigma^2_{Ga} = 0.5, \sigma^2_{Gd} = 0, \sigma^2_e = 0.4, q_1 = P_A = 0.5$, *which are the same as those of Table 2 of Abecasis, Cookson, and Cardon (2000). The power of* χ^2_{qtl} *is taken from Table 2 of Abecasis, Cookson, and Cardon (2000).*

D'	0.00	0.25	0.50	0.75	1.00
Graph A: small 3-generation pedigree					
χ^2_{qtl}	0.6	7.8	33.5	76.5	98.4
$F_{A,a}$	1.0	18.2	77.7	99.3	100.0
Graph B: large 3-generation pedigree					
χ^2_{qtl}	0.6	16.4	73.3	99.1	100.0
$F_{A,a}$	1.0	39.3	97.9	100.0	100.0

2.4 Discussion

This chapter focuses on constructing models which utilize multi-generational pedigrees for LD mapping of QTL. It is an extension of previous works (Fan and Jung (2003); Fan and Xiong (2003)), in which variance component models are constructed for high resolution joint linkage and LD mapping of QTLs. The proposed method has higher power than the "AbAw" approach. These finding are consistent with previous observations based on sib-pair data by Fan and Jung (2003). Our power comparisons confirm that large pedigrees may contain more LD information than small pedigrees.

Our method decomposes the association into the summation of additive and dominant effects. Additionally, when the dominant effect is not significantly different from 0, then only the additive effect is modeled. The "AbAw" approach, on the other hand, decomposes the association into the summation of between-family (b) and within-family (w) components. It is possible that this difference explains the increased power observed for our method.

Population data can contain LD information, while pedigree data may contain both linkage and LD information. Thus, it is interesting to combine pedigree data with population data for fine association studies. The methods developed in this chapter allow the mapping of QTLs in a unified linkage and LD analysis using a combination of pedigree and population data. Linkage analysis can be used to localize genetic traits to a broad region, and is less sensitive to population structures than LD mapping. LD mapping, on the other hand, is appropriate for high resolution mapping and sensitive to population admixtures. In an examination of real data, it is sensible to first perform linkage analysis using pedigree data to identify regions which demonstrate linkage to the trait of interest. Sparse genetic maps are appropriate for this step in the analysis, see for example Broman et al. (1998) or Kong et al. (2002).

Then, a combined analysis using both population and pedigree data can be performed to take advantage of both linkage and LD information for high resolution mapping of genetic traits. This method can provide the high resolution characteristic of LD mapping, while simultaneously providing additional protection against false positives by using the prior linkage evidences based on a dense genetic map.

CHAPTER III

GENE-ENVIRONMENT INTERACTIONS WITH MISSING GENETIC INFORMATION

3.1 Introduction

Many important human diseases, including most cancers, have both genetic and environmental risk factors. Interactions between these factors can sometimes be more important than the values of the individual factors in determining the probability of developing disease (Yang and Khoury (1997)). Understanding the complex relationships between genetics and the environment allows medical interventions to be custom tailored to an individual's genetic makeup. Although there are many possible study designs used to investigate these questions, case-control studies are often the most practical, due to their ability to be used for both rare and late-onset diseases. Additionally, since part or all of the genetic information may be missing for many of the subjects in the study, incorporating information from subjects with missing genetic information may increase the precision of our estimates. Thus, we develop a method for estimating gene-environment interactions for case control studies with missing genetic information.

The motivation for the development of this method came from a study which examines the development of ovarian cancer in a population of Israeli women. One goal of this study is to investigate the interaction between either a BRCA1 or BRCA2 mutation and both oral contraceptive usage and parity. Additionally, many other covariates, including age, family history of breast and ovarian cancer, personal history of breast cancer, history of gynecological surgery, and ethnicity, are measured as they are believed to also effect the risk of developing the cancer. Unfortunately, since the

mutations of interest are also important in breast cancer, many of the subjects in the study have declined to provide samples for genetic testing.

Unknown genetic information also occurs when haplotypes are the genetic component of interest, but only genotypes are measured. This is a frequent occurrence, as methods to determine haplotype directly are much more costly than methods to assess genotype alone. Many previous works have examined methods for inferring the haplotypes when they are not known. The first such method is due to Clark (1990)and attempts to explain all observed genotypes using the minimum number of haplotypes. The method is simple to implement, often performs well in practice, and has been used successfully in many applications. More recently, methods have been proposed based upon maximum likelihood estimation (Excoffier and Slatkin (1995), and Fallin and Schork (2000)), or upon Bayesian methodology (Niu et al. (2002) and Thomas et al. (2001)). Other methods attempt to analyze the data including the uncertainty in haplotype phase (Schaid et al. (2002), Wallenstein et al. (1998), and Zhao et al. (2003)). Our method does not infer the haplotypes directly, but rather estimates the probability of occurrence of each possible haplotype and incorporates all of the information into the likelihood.

The method that we develop finds the semiparametric maximum likelihood estimates of the parameters of a logistic regression in order to estimate the probability of disease. The distribution of the genotype in the population is assumed to be discrete. Furthermore, as little is known about the joint distribution of the environmental factors, they are treated nonparametrically. Prentice and Pyke (1979) show that if the distribution of the environment is left completely unspecified then the logistic parameters may be estimated as if the study was performed prospectively. In our method, however, we make the additional assumption that the distributions of the genes and the environment are marginally independent, conditional upon any variables which

describe any stratification in the population, see Umbach and Weinberg (1997) and Satten et al. (2001). This follows the work of Chatterjee and Carroll, which shows that, under this assumption, there are large gains in the efficiency of estimation. Adopting this framework, we calculate the complete likelihood for the data under the case-control sampling design. Subsequently, the profile likelihood is used to calculate the semiparametric maximum likelihood estimates for the parameters of interest without calculating the distribution of the environment.

This chapter is organized as follows. First, in Section 3.2.1, we develop the methodology and asymptotic distributional results for a homogeneous population, under the assumption that the distribution of the gene and environment are marginally independent. Next, in Section 3.2.2 we describe a modification to the method that can be used when the probability of disease in the population is known. This modification is shown to decrease the variability of the parameter estimates when the probability of disease can be well estimated for the population. In Section 3.3.1 we present the results of simulation studies that show that parameter estimates are unbiased and have approximately the claimed asymptotic standard errors. Lastly, Section 3.4 contains some concluding remarks. Note that all proofs are provided in Appendix B.

3.2 The Method

3.2.1 Homogeneous Populations with Disease Probability Unknown

Consider a case-control study with n_0 controls and n_1 cases. Let D_i denote the disease state of individual i, X_i denote the vector of their environmental covariates, G_i denote their true genotype, G_i^* denote values of their genotype that are measured and let Δ_i be a variable whose values indicate what genetic information is measured. For

example, in a haplotype study, we could have

$$\Delta = \begin{cases} 1 & \text{if no genetic information is measured;} \\ 2 & \text{if unphased genotype is measured;} \\ 3 & \text{if diplotype is measured.} \end{cases}$$

Assume that the distribution of the genotype and the environment are marginally independent and that what type of genetic information is measured does not depend upon the individual's underlying genetic makeup. Mathematically, the second of these two assumptions can be written as $\text{pr}(\Delta|D,X,G) = \text{pr}(\Delta|D,X) = \pi(\Delta|D,X,\xi)$. Furthermore, treat the distribution of the genotypes as discrete with $\text{pr}(G = g_j) = h(g_j,\theta)$ where $h(\cdot)$ is a known function and θ is a vector of parameters. Define $\mathcal{G}_i = \{g_j : g_j \text{ is consistent with } G_i^*\}$ to be the set of all possible genotypes which are consistent with the genetic information observed and assume that the probability of disease depends only upon the true genotype of an individual, not upon the portion of the genotype that is measured. This assumption can be stated as $\text{pr}(D|X,G,G^*) = \text{pr}(D|X,G)$. Also, observe that $\text{pr}(G,G^*) = \text{pr}(G)$ for all $G \in \mathcal{G}$. Then,

$$\begin{aligned} \text{pr}(D|X,G^*) &= \sum_j \text{pr}(D|X,G=g_j,G^*)\text{pr}(G=g_j|X,G^*) \\ &= \sum_j \text{pr}(D|X,G=g_j)\text{pr}(G=g_j|G^*) \\ &= \sum_{g_j \in \mathcal{G}} \frac{\text{pr}(D|X,G=g_j)h(g_j,\theta)}{\text{pr}(G^*)}. \end{aligned}$$

Now, treat the distribution of the environmental covariates as a discrete distribution having mass at every observed value; thus $\text{pr}(X = x_k) = \zeta(x_k)$. Then, the

retrospective log likelihood has the form

$$\begin{aligned}\log L(\beta_0,\beta_1,\theta,\zeta,\pi) &= \sum_{i=1}^{n}\left[\log\left\{\frac{\mathrm{pr}(D_i|X_i,G_i^*)\mathrm{pr}(G_i^*)\mathrm{pr}(X_i)\mathrm{pr}(\Delta_i|D_i,X_i)}{\mathrm{pr}(D_i)}\right\}\right] \\ &= \sum_{i=1}^{n}\left[\log\left\{\sum_{g_j\in\mathcal{G}_i}\mathrm{pr}(D_i|X_i,g_j)h(g_j,\theta)\right\}\right. \\ &\quad \left.+\log\{\zeta(X_i)\}+\log\{\pi(\Delta_i|D_i,X_i)\}-\log\{\mathrm{pr}(D_i)\}\right].\end{aligned}$$

Notice that the likelihood consists of terms which depend upon $\beta_0,\beta_1,\theta,\zeta$ and those which depend upon $\pi(\cdot)$ and thus maximization with respect to $\beta_0,\beta_1,\theta,\zeta$ can be completed independent of $\pi(\cdot)$. Now, define $\kappa=\beta_0+\log[\{n_1\mathrm{pr}(D=0)\}/\{n_0\mathrm{pr}(D=1)\}]$, $\Omega=(\beta_0,\beta_1,\theta,\kappa)$, $H(x)=\{1+\exp(-x)\}^{-1}$, and $S(D,X,G,\Omega)=h(G,\theta)\exp[D\{\kappa+m(X,G,\beta_1)\}][1-H\{\beta_0+m(X,G,\beta_1)\}]$.

Lemma 1: The profile likelihood function has the form

$$\log\{L(\Omega)\}=\ell(\Omega) = \sum_{i=1}^{n}\left[\log\sum_{g_j\in\mathcal{G}_i}S(D_i,X_i,g_j,\Omega)-\log\sum_{d,j}S(d,X_i,g_j,\Omega)\right].$$

Define $I=-(1/n)\mathrm{E}\{\partial^2\ell(\Omega)/\partial\Omega\partial\Omega^T\}$ and $\Lambda=\sum_d(n_d/n)\mathrm{E}\{\Psi(\Delta,D,X,G^*,\Omega)|D=d\}\times[\mathrm{E}\{\Psi(\Delta,D,X,G^*,\Omega)|D=d\}]^T$, where

$$\Psi(\Delta,D,X,G^*,\Omega) = \frac{\sum_{g_j\in\mathcal{G}_i}S_\Omega(D_i,X_i,g_j,\Omega)}{\sum_{g_j\in\mathcal{G}_i}S(D_i,X_i,g_j,\Omega)}-\frac{\sum_{d,j}S_\Omega(d,X_i,g_j,\Omega)}{\sum_{d,j}S(d,X_i,g_j,\Omega)}.$$

Then, we have the following lemma.

Lemma 2: Under suitable regularity conditions, and for fixed n_0/n, the solutions which maximize the profile likelihood satisfy $n^{1/2}(\widehat{\Omega}-\Omega_0)\Rightarrow\mathrm{N}(0,\Sigma)$ where

$$\Sigma = I^{-1}-I^{-1}\Lambda I^{-1}.$$

3.2.2 Probability of Disease Known

In many human diseases, the probability of occurrence of the disease in the general population is known. The estimation procedure the we have developed can easily be modified to incorporate the known value. First, define $\Pr(D = 1) = p$, $\nu = \log[\{n_1(1-p)\}/(n_0 p)]$. Then, $\kappa = \beta_0 + \nu$, $\Omega = (\beta_0, \beta_1, \theta)$ and $S(D, X, G, \Omega) = h(G, \theta)\exp[D\{\beta_0 + \nu + m(X, G, \beta_1)\}][1 - H\{\beta_0 + m(X, G, \beta_1)\}]$. In this setting, we can again find the values of the parameters that maximize the function

$$\ell(\Omega) \;=\; \sum_{i=1}^{n}\left[\log\left\{\sum_{g_j\in\mathcal{G}_i} S(D_i, X_i, g_j, \Omega)\right\} - \log\left\{\sum_{d,j} S(d, X_i, g_j, \Omega)\right\}\right].$$

Similar to the case when the probability of disease in the population is not known, define $I = -(1/n)\mathrm{E}\{\partial^2\ell(\Omega)/\partial\Omega\partial\Omega^T\}$ and $\Lambda = \sum_d (n_d/n)\mathrm{E}\{\Psi(\Delta, D, X, G^*, \Omega)|D = d\} \times [\mathrm{E}\{\Psi(\Delta, D, X, G^*, \Omega)|D = d\}]^T$, where

$$\Psi(\Delta, D, X, G^*, \Omega) \;=\; \frac{\sum_{g_j\in\mathcal{G}_i} S_\Omega(D_i, X_i, g_j, \Omega)}{\sum_{g_j\in\mathcal{G}_i} S(D_i, X_i, g_j, \Omega)} - \frac{\sum_{d,j} S_\Omega(d, X_i, g_j, \Omega)}{\sum_{d,j} S(d, X_i, g_j, \Omega)}.$$

Then, under suitable regularity conditions, and for fixed n_0/n, the solutions which maximize $\ell(\Omega)$ satisfy $n^{1/2}(\widehat{\Omega} - \Omega_0) \Rightarrow \mathrm{N}(0, \Sigma)$ where $\Sigma = I^{-1} - I^{-1}\Lambda I^{-1}$.

3.3 Simulation Study

3.3.1 Goals

In this section we present the results of a comprehensive simulation study. The purpose of this study was to investigate different properties of the estimates, including determining the magnitude and direction of any biases, validating the estimated and asymptotic variances, and assessing the effect of knowing the probability of disease. Additionally, we wish to determine the effects of the frequency of a mutation by comparing a rare mutation with a common mutation. Furthermore, we consider the problem of estimating the interaction between a haplotype and the environment,

however, we assume that the investigators only measure the genotype. Thus, some of the genetic information is missing for a fraction of the individuals in the study.

For all of the simulations, the maximization is performed using Fisher's method of scoring. The methodology is quite stable when the probability of disease is known, however, when the probability of disease is not known, the maximization may be unstable if the starting value for β_0 is far from the maximum value. For this reason, for the case when the probability of disease is not known, simulations are performed using a grid search to identify the value of β_0 which provides the largest value of the likelihood when all other parameters are maximized.

3.3.2 Simulation Design and Results

To determine the effect of knowing the probability of disease as compared to estimating it from the data, we conducted a simulation study. In this study we simulated 1000 data sets, each containing 1000 cases and 1000 controls from the population. Let the environmental covariate be distributed as min{exp(X),10)} where $X \sim \mathrm{N}(0,1)$, and consider the genotype at two loci with two alleles each. Then, when the first locus has alleles A and a, and the second locus has alleles B and b, we have four possible haplotypes, (AB,Ab,aB,ab). In the population, let these haplotypes have probabilities $(.4, .3, .2, .1)$ for the "common" haplotype or $(.425, .325, .225, .025)$ for the "rare" haplotype. Let the ab haplotype be associated with an increased disease risk, as compared to the other haplotypes, and combine the haplotypes under Hardy-Weinberg equilibrium to form genotypes. Now, let the probability of disease follow

$$\mathrm{logit}\{\mathrm{pr}(D=1|G,X)\} \;=\; \beta_0 + \beta_G(\text{num of ab}) + \beta_X X + \beta_{GX} X(\text{num of ab}),$$

where $\beta_0 = -3.5, -3.25$ for the common and rare haplotypes, $\beta_G = .26$, $\beta_X = .1$, and $\beta_{GX} = .35$. For each simulated data set, the analysis is performed first

considering that the probability of disease is known and second that it is unknown. Additionally, estimated values of the variances of the parameters are calculated for each data set under each scenario by estimating the asymptotic variance using the observed data values. These results are then compared to both the observed variances from simulation and the true asymptotic variances.

The results of the simulations for the frequency are presented in Table 2. For all cases, the observed biases are quite small. Additionally, the estimated standard errors correspond closely to the observed errors from simulation. This suggests that the approximate values obtained for a single data set can be used to form confidence intervals and perform tests using the asymptotic distribution. The results also indicate an important trend when the analysis is performed assuming that the probability of disease is known. The simulations suggest that there is a decrease in the standard errors of the parameter estimates for all parameters, when the probability of disease is known. For most human diseases, the incidence of disease is either known, or can be very well estimated, and thus this information should be incorporated into the data analysis whenever possible.

3.4 Discussion

One area of particular interest in genetic epidemiology today is understanding the interactions between genetic factors and the environment. The ability to study these sorts of interactions with a case-control study is particularly valuable, as these studies are flexible and appropriate for many human diseases of interest.

The methods developed in this chapter allow the estimation of the interaction between the gene and the environment for case-control studies, even in the event that some study subjects have part or all of their genetic information missing. This flexibility allows the method to incorporate information about environmental factors

Table 2. The results of a simulation for a homogeneous population with 1000 replications for a case-control study with 1000 cases and 1000 controls. Values presented include the observed bias and observed, estimated, and asymptotic standard errors for each of the parameters of interest. Additionally, the simulations assess the effects of knowing the probability of disease for both a common and a rare haplotype.

Haplotype Case	Pr(D=1)	Beta	Bias	Observed Standard Error	Estimated Standard Error	Theoretical Standard Error
Common	Known	β_G	-0.0029	0.1329	0.1343	0.1327
		β_X	-0.0018	0.0260	0.0261	0.0267
		β_{GX}	0.0055	0.0456	0.0444	0.0429
	Unknown	β_G	-0.042	0.1574	0.1577	0.1587
		β_X	-0.0013	0.0292	0.0287	0.0298
		β_{GX}	0.0084	0.0594	0.0576	0.0568
Rare	Known	β_G	0.0003	0.2652	0.2605	0.2626
		β_X	0.0010	0.0243	0.0244	0.0247
		β_{GX}	0.0043	0.0884	0.0859	0.0860
	Unknown	β_G	-0.0181	0.3073	0.3124	0.3195
		β_X	0.0006	0.0244	0.0251	0.0256
		β_{GX}	0.0172	0.1143	0.1171	0.1158

into our analysis even when the individuals have no genetic information measured. The advantage of incorporating such individuals is that their inclusion decreases the standard errors of all parameters in the study, even those pertaining to genetic effects. Additionally, this method can be used to investigate the relationship between haplotypes and the environment. Furthermore, the method does not require that haplotypes are measured for any individuals; in fact, for some measured genotypes, the unique haplotype may be inferred.

Another advantage to this procedure is that the distribution of the missing values may depend upon both the disease state and the environmental values. This flexibility permits the method to be used in cases where control subjects are much more likely to have missing information than case subjects. This makes the method very valuable in

many human studies where control subjects do not wish to be genotyped based upon their confidentiality concerns. Additionally, this flexibility also allows the scientist to select subjects to genotype based upon their disease state and environmental values in the event that costs may be decreased by only genotyping a portion of the entire study population.

Finally, the assumption of marginal independence between the gene and the environment is satisfied for many environmental factors of interest. When this assumption is valid, our proposed method produces parameter estimates having smaller standard errors than parameter estimates from traditional analyses without this assumption. Additionally, data collected about the distribution of the environment and disease state, even when the genotype is unknown, provides a significant improvement in the precision of parameter estimates. In the event that the probability of disease is known, our method incorporates this information, again decreasing the standard errors of our estimates. These results are not true for a traditional analysis, in which knowing the probability of disease does not effect the standard errors of the parameter estimates.

CHAPTER IV

GENE-ENVIRONMENT INTERACTIONS WITH POPULATION STRATIFICATION

4.1 Introduction

Many naturally occurring populations are stratified, meaning that the population is composed of a variety of homogeneous sub-populations having different backgrounds, and thus also different genetic makeup, disease risk, and/or environmental exposures. These sub-populations are often racial or ethnic, and some members of the population may be a mixture or two or more of the sub-populations. In the presence of a population with this type of structure, the assumption of independence between genetic factors and the environment may be violated. However, within a particular sub-population, the assumption of independence may still hold. When we can measure covariates which determine the identity of each individual's sub-population, we can then modify our method to apply in this case.

This chapter is organized as follows. In Section 4.2, we discuss modifications to the method proposed in Chapter III that are needed to account for population stratification. Also discussed are the necessary modifications that allow these methods to be used for frequency matched case-control studies. We present a simulation study to evaluate the performance of these methods in Section 4.3, while in Section 4.4 we provide a few concluding remarks. Proofs may be found in Appendix C.

4.2 Method

4.2.1 Stratified Populations

First, consider each sub-population to be a strata, and assume that covariates which determine the value of the strata are measured. Assume that the distribution of the genotype, G, and the environment, X, are independent given the strata, S. We can then perform the analysis in a similar method to that developed in Chapter III. Furthermore, assume that the joint distribution of the gene, environment, and stratification variables can be written as the joint distribution of environment and stratification variables times the conditional distribution of the genotype given the stratification variables. Then, if the number of strata is small, $h(G_j, \theta)$ can be replaced by $h(g_j, \theta; S)$, the conditional distribution of genotype given strata, where the $h_S(\cdot)$ are known functions and θ is a vector of parameters. Also, the joint distribution of the environment and the strata may be treated nonparametrically by replacing $\zeta_k = \text{pr}(X = x_k)$ by $\zeta_{k,m} = \text{pr}(X = x_k, S = s_m)$. In the event that the number of strata grows with the sample size, the above procedure may still be used if a parametric model with fixed $\dim(\theta)$ is assumed for $\text{pr}(G = g_j | S = s)$. In either of these cases, the results from Chapter III again follow with the appropriate substitutions for $h(g_j, \theta)$ and ζ_k. These results can be derived using calculations similar to those in Chatterjee and Carroll.

4.2.2 Frequency Matched Case-Control Studies

This methodology may also be extended to account for the frequency matched case-control study design. First, let $W = w_m, m = 1, ..., M$ denote the M strata used for matching. Also, let $W^S = S^W$ be the elements of the environment that are used for both matching and population stratification, $W^{\overline{S}}$ be the matching variables that are not involved in population stratification, and $S^{\overline{W}}$ be the population stratification

variables that are not used for matching. Then, assume that $\text{pr}(G = g_j|S = s)$ is independent of $(X, W^{\overline{S}})$ and that $\text{pr}(D = 1|G, X, S, W) = H\{\beta_{0,W} + m(G, X, S, W, \beta_1)\}$. Then the log-likelihood for the matched case-control study is

$$\begin{aligned}
\log L_{\text{mcc}} &= \sum_{i=1}^{n} \left[\log\left\{\text{pr}(G_i^*, \Delta_i, X_i, S_i^{\overline{W}}|D_i, W_i)\right\}\right] \\
&= \sum_{i=1}^{n} \left[\log\left\{\sum_{g_j \in \mathcal{G}_i} \text{pr}(D_i|g_j, X_i, S_i, W_i)h(g_j|S_i)\right\} + \log\{\text{pr}(X_i, S_i, W_i)\}\right. \\
&\quad \left. + \log\{\text{pr}(\Delta_i|D_i, G_i^*, X_i, S_i, W_i)\} - \log\{\text{pr}(D_i, W_i)\}\right].
\end{aligned}$$

Lemma 3: Under the frequency matched case-control study design, the data may be analyzed as though an unmatched study was performed if either (1) the probability of disease is known for each value of the matching variable or (2) if the probability of disease is very small for all possible gene-environment combinations.

4.3 Simulation Study Design and Results

To investigate the effects of stratification in the population, a simulation study for such a population was performed. First, 1000 cases and 1000 controls were sampled from the population. Two strata are present with 60% of the population in the first stratum and 40% of the population in the second stratum. The environmental covariate is distributed as min{exp(X),10} where $X \sim \text{N}(\mu, 1)$, where $\mu = 0, .67$ for the two strata. These values are selected such that the 75th percentile of the first population occurs at the median for the second population. We again consider four possible haplotypes, (AB,Ab,aB,ab), with probabilities (.4, .3, .2, .1) and (.35, .275, .175, .2) for the "common" haplotype and (.425, .325, .225, .025) and (.4, .325, .225, .05) for the "rare" haplotype, for the two strata, respectively. In both cases, the second stratum has double the occurrence of the disease-associated haplotype. Additionally, the last

haplotype is associated with increased disease risk and the haplotypes are combined under Hardy-Weinberg equilibrium. The disease probability then follows

$$\begin{aligned}\text{logit}\{\text{pr}(D=1|G,X,S)\} &= \beta_0+\beta_G(\text{num of ab})+\beta_X X+\beta_S S \\ &\quad +\beta_{GX}X(\text{num of ab})+\beta_{GS}S(\text{num of ab}),\end{aligned}$$

where $\beta_0 = -3.5, -3.25$, for the common and rare haplotypes, and $\beta_G = .26$, $\beta_X = .1$, $\beta_S = -.2$, $\beta_{GX} = .35$, and $\beta_{GS} = -.25$.

Table 3. The results of a simulation for a stratified population with 1000 replications for a case-control study with 1000 cases and 1000 controls. The underlying population is composed of two strata which differ in their disease risk, genotype distribution, and gene-environment interaction. Values presented include the bias and the observed, estimated, and asymptotic standard error for each of the parameters of interest. Additionally, the simulations examine the effects of the frequency of the hayplotype on these parameters.

Haplotype Case	Beta	Bias	Observed Standard Error	Estimated Standard Error	Theoretical Standard Error
Common	β_G	-0.0071	0.1662	0.1645	0.1640
	β_X	-0.0020	0.0248	0.0246	0.0246
	β_{GX}	0.0054	0.0358	0.0357	0.0350
	β_S	0.0056	0.1312	0.1349	0.1345
	β_{GS}	-0.0083	0.2107	0.2106	0.2074
Rare	β_G	-0.0129	0.3002	0.3148	0.3069
	β_X	¡ 0.0001	0.0210	0.0207	0.0203
	β_{GX}	0.0065	0.0626	0.0623	0.0604
	β_S	0.0019	0.1034	0.1070	0.1072
	β_{GS}	0.0044	0.4026	0.4004	0.3935

The results presented in Table 3 indicate that the trends observed in a homogeneous population hold also in the presence of population stratification. First, the biases are small in magnitude, and do not display a consistent direction. Furthermore, the estimated standard errors are similar to those observed from simulation. However, with a stratified population, it appears that a larger sample size would be needed in

order for the estimated standard error to achieve the same quality of approximation for the asymptotic standard errors as is seen in the case of a homogeneous population.

4.4 Discussion

Population stratification is one of the largest hurdles faced by case-control studies. Left uncorrected, it can severely bias estimates of the parameters of interest. However, when variables are measured which define the strata within the population, the method from Chapter III may again be used to find unbiased estimates. Estimates of the standard errors of the semiparametric maximum likelihood estimates are also quite close to the true values. Thus, our method provides a means to form confidence intervals or perform tests of hypotheses about the model parameters. Finally, this test, in the presence of population stratification, shares many of the other benefits discussed for the method in Chapter III.

CHAPTER V

GENE-ENVIRONMENT INTERACTIONS WITH GENOTYPE MISCLASSIFICATION

5.1 Introduction

Recently, some attention has focused upon the problems that arise when genotypes are misclassified. For example, Wong et al. (2004), study the effects of genotype misclassification on the gene-environment interaction parameter in a linear regression. Previous methods for analyzing the interactions between genetic and environmental factors in a logistic regression to estimate the probability of disease do not address these types of concerns. However, the method proposed by Chatterjee and Carroll can be extended to accommodate this additional complication.

This chapter is organized as follows. In Section 5.2, we develop a method to handle genotype misclassification in a study of gene-environment interactions similar to that in previous chapters. The method simplifies to the method of Chapter III in the event that there is no misclassification of the genotype. In Section 5.3 we discuss a simulation study used to examine the effect that the rate of misclassification has on the method. The simulation study also examines the effect of ignoring misclassification error when it is present. Lastly, in Section 5.4, we provide concluding remarks. All proofs may be found in Appendix D.

5.2 The Method

5.2.1 Misclassification Probabilities Known

Consider a case-control study with n_0 controls and n_1 cases. Let D_i denote the disease state of individual i, X_i denote the vector of their environmental covariates,

G_i denote their true genotype, and let Δ_i be a variable whose values indicate what genetic information is measured. For example, for a haplotype study, we could have

$$\Delta = \begin{cases} 1 & \text{if no genetic information is measured;} \\ 2 & \text{if unphased genotype is measured;} \\ 3 & \text{if haplotype is measured.} \end{cases}$$

Assume also that the type of genetic information measured does not depend upon the true genotype; that is, $\text{pr}(\Delta|D,X,G) = \text{pr}(\Delta|D,X) = \pi(\Delta|D,X,\xi)$. For each value of Δ, the space of possible genotypes is divided into L_k partitions, denote these by $\zeta_{k1},...,\zeta_{kl}$. Then for each individual, we observe one of the partitions, $\mathcal{M}_i$, however the true (unobserved) partition is $\mathcal{G}_i$. That is, $G_i \in \mathcal{G}_i$ but sometimes $G_i \notin \mathcal{M}_i$. Assume that the probability of seeing a particular partion does not depend upon the environment, $\Pr(\mathcal{M}_i = \zeta_{k\ell}|D,X,\mathcal{G}_i,\Delta_i) = \Pr(\mathcal{M}_i = \zeta_{k\ell}|D,\mathcal{G}_i,\Delta_i) = \pi_{\mathcal{M}k}(\ell|\mathcal{G}_i,D,\eta)$, where η is a vector of parameters that specify the distribution. Now, assume the the distribution of the genetic factors and the environmental factors are independent and treat the distribution of the genotypes as discrete with $\text{pr}(G = g_j) = h(g_j,\theta)$ where $h(\cdot)$ is a known function and θ is a vector of parameters.

Next, define $\kappa = \beta_0 + \log[\{n_1\text{pr}(D=0)\}/\{n_0\text{pr}(D=1)\}]$, $\Omega = (\beta_0,\beta_1,\theta,\kappa)$, $H(x) = \{1+\exp(-x)\}^{-1}$, and $S(D,X,G,\Omega) = h(G,\theta)\exp[D\{\kappa + m(X,G,\beta_1)\}][1 - H\{\beta_0 + m(X,G,\beta_1)\}]$. Further, define the pseudo-likelihood for a single observation and a particular (known) value of G as

$$L(\Omega) = \frac{S(D,X,G,\Omega)}{\sum_{d,j} S(d,X,g_j,\Omega)},$$

so that the pseudo-likelihood for an individual has the form

$$\prod_{k=1}^{K}\left(\frac{\pi(k|D,X,\xi)\sum_{\ell,g\in\zeta_{k\ell}} S(D,X,g,\Omega)\prod_s\{\pi_{\mathcal{M}k}(s|\zeta_{kl},D,\eta)\}^{I(\mathcal{M}=\zeta_{ks})}}{\sum_{d,j} S(d,X_i,g_j,\Omega)}\right)^{I(\Delta=k)}.$$

We can use the above equation to derive the following lemma.

Lemma 4: The log of the pseudo-likelihood function, denoted $\ell(\Omega)$ has the form

$$\begin{aligned}\ell(\Omega) &= \sum_{i=1}^{n}\left\{\sum_{k=1}^{K} I(\Delta_i = k)\log\left(\sum_{\ell=1}^{L_k}\left[\sum_{g\in\zeta_{k\ell}} S(D_i, X_i, g, \Omega)\right.\right.\right. \\ &\quad \left.\left.\left.\times \prod_{s=1}^{L_k}\{\pi_{\mathcal{M}_i k}(s|\zeta_{kl}, D_i, \eta)\}^{I(\mathcal{M}_i=\zeta_{ks})}\right]\right) - \log\{\sum_{d,j} S(d, X_i, g_j, \Omega)\}\right\}.\end{aligned}$$

Define

$$I = -\frac{1}{n}\mathrm{E}\left[\frac{\partial^2}{\partial\Omega\partial\Omega^T}\log\{L(\Omega)\}\right]$$

and

$$\Lambda = \sum_d \frac{n_d}{n}\mathrm{E}\{\Psi(\Delta, D, X, \mathcal{M}, \Omega)|D = d\}\left[\mathrm{E}\{\Psi(\Delta, D, X, \mathcal{M}, \Omega)|D = d\}\right]^T$$

where

$$\begin{aligned}\Psi(\Delta, D, X, \mathcal{M}, \Omega) &= -\frac{\sum_{d,j} S_\Omega(d, X, g_j, \Omega)}{\sum_{d,j} S(d, X, g_j, \Omega)} \\ &\quad + \sum_{k=1}^{K} I(\Delta = k)\frac{\sum_{\ell=1}^{L_k}\left[\sum_{g\in\zeta_{k\ell}} S_\Omega(D, X, g, \Omega)\prod_{s=1}^{L_k}\{\pi_{\mathcal{M}k}(s|\zeta_{kl}, D, \eta)\}^{I(\mathcal{M}=\zeta_{ks})}\right]}{\sum_{\ell=1}^{L_k}\left[\sum_{g\in\zeta_{k\ell}} S(D, X, g, \Omega)\prod_{s=1}^{L_k}\{\pi_{\mathcal{M}k}(s|\zeta_{kl}, D, \eta)\}^{I(\mathcal{M}=\zeta_{ks})}\right]}.\end{aligned}$$

Using the above definitions and a repeated application of the central limit theorem, we have the following lemma. **Lemma 5:** Under suitable regularity conditions, and for fixed n_0/n, the solutions to the score satisfy $\sqrt{n}(\widehat{\Omega} - \Omega_0) \longrightarrow_d \mathrm{N}(0, \Sigma)$ where

$$\Sigma = I^{-1} - I^{-1}\Lambda I^{-1}$$

5.2.2 Misclassification Probabilities Unknown

The previous section deals with the situation where the probability of misclassification is either known or estimated from an external data source. However, sometimes it is of interest to consider an internal validation study. In this case, we will need to also

estimate the parameters, η, along with $\beta_0, \beta_1, \theta, \kappa$. Fortunately, we can recalculate the forms above, using $\Omega = (\beta_0, \beta_1, \theta, \kappa, \eta)$ and all of the results continue to hold, with the appropriate changes in the definition of Ω.

5.3 Simulations

To better understand the effect of genetic misclassification, we perform a simulation study to determine that magnitude of any biases and to evaluate the appropriateness of the asymptotic results. Toward this end, we are considering a model for misclassification similar to that used by Wong, et al (2004). We consider a single genetic locus with two possible alleles, A and a and assume that misclassification of one of the two alleles in an individual is independent of the classification of the other. Additionally, assume that the two alleles are misclassified at the same rate.

We consider a population with 1000 cases and 1000 controls in which all of the individuals have their genotype at one locus measured, and possibly misclassified. Additionally, we let 200 cases and 200 controls be part of a validation study and thus, these individuals are genotyped twice. Now, let the true probability that an allele is misclassified be $p = .05, .1$ and let the frequency of the A allele in the population be $p_A = .7, .9$. Assume that the double recessive genotype has increased disease risk and allow the environmental variable to be distributed as min{exp(X),10)} where $X \sim \mathrm{N}(0, 1)$. Next, let the probability of disease follow

$$\mathrm{logit}\{\mathrm{pr}(D = 1|G, X)\} \quad = \quad \beta_0 + \beta_G I(aa) + \beta_X X + \beta_{GX} X I(aa),$$

where $\beta_0 = -3.5, -3.25$ for $p_A = 0.7, 0.9$, $\beta_G = .26$, $\beta_X = .1$, and $\beta_{GX} = .35$. For each simulated data set, the analysis is performed first accounting for possible misclassification and again ignoring misclassification and using only the original genotype. Furthermore, estimated values of the variances of the parameters are calculated for

each data set in each scenario by estimating the asymptotic variance using the observed data values. These results are compared to both the observed variances from simulation and the true asymptotic variances.

Table 4. The results of a simulation for genotype misclassification with 1000 replications for a case-control study with 1000 cases and 1000 controls. Values presented include the observed bias and the standard error for each of the parameters calculated from the simulated values, along with the average estimated standard error and asymptotic standard error. Additionally, the results compare two rates of misclassification, as well as two frequencies of the A allele.

p_{mis}	p_A	Beta	Bias	Bias Ignoring Misclass.	Observed Standard Error	Estimated Standard Error	Theoretical Standard Error
0.05	0.70	β_G	-0.0244	-0.0742	0.1645	0.1721	0.1620
		β_X	-0.0002	0.0113	0.0251	0.0244	0.0235
		β_{GX}	0.0069	-0.0546	0.0586	0.0565	0.0535
0.05	0.90	β_G	-0.0322	-0.1563	0.6231	0.5571	0.4069
		β_X	0.0005	0.0020	0.0236	0.0238	0.0221
		β_{GX}	0.0133	-0.1418	0.2140	0.1858	0.1366
0.10	0.70	β_G	-0.0062	-0.1078	0.2009	0.1999	0.1769
		β_X	-0.0013	0.0195	0.0255	0.0254	0.0244
		β_{GX}	0.0037	-0.1083	0.0621	0.0637	0.0557
0.10	0.90	β_G	-0.1121	-.2026	1.0517	0.9044	0.4375
		β_X	-0.0002	0.0021	0.0245	0.024	0.0230
		β_{GX}	0.0570	-.2199	0.4267	0.3025	0.1610

The results of this simulation study are found in Table 4 and reveal several important features of this method. First, the estimates provided by our method appear to be unbiased, and the estimated standard errors provide good fit to the standard errors observed from simulation, for a common allele or low rate of misclassification. Second, it appears that as the amount of misclassification increases, so do the standard errors for the parameters of interest. Next, as the amount of misclassification increases, so do the biases that are associated with ignoring misclassification. Moreover, these biases are quite severe, even when the rate of misclassification is $p_{mis} = .05$.

Finally, when the frequency of one allele becomes small and the misclassification rate becomes large, even our corrected method performs poorly for this sample size.

5.4 Discussion

The problem of misclassification of genotypes has received increasing interest in recent years. Misclassification, when ignored, can lead to biases in parameter estimates. Thus, we develop a method which can incorporate the possible genotype misclassification into a pseudo-likelihood analysis to estimate gene-environment interactions.

This method has several benefits. First, it provides unbiased estimates of the effects of genotype, environment, and the gene-environment interaction for cases of moderate misclassification. Furthermore, the method that we develop allows estimation the standard errors of these estimates for a single data set; this estimation procedure provides good approximations to the true standard errors. Lastly, the asymptotic distributional results allow the construction of confidence intervals or hypothesis tests for the parameters of interest.

The importance of methods which account for genotype misclassification is also demonstrated in this chapter. As previously noted, when misclassification is present and unaccounted for, it can seriously bias the parameter estimates in a logistic regression. This could easily lead to parameters being declared significant when in fact they are not, or to interactions being declared insignificant, when if fact they are important in determining the probability of disease.

CHAPTER VI

ANALYSIS OF ISRAELI OVARIAN CANCER STUDY DATA

6.1 Introduction

In this chapter, we analyzed a frequency matched case-control study of ovarian cancer in Israeli women in order to investigate the interaction between the BRCA1/BRCA2 mutations and 1) oral contraceptive use or 2) parity in determining the probability of disease. For each of the women in the study, investigators measure disease state, presence of BRCA1/BRCA2 mutation, parity, length of oral contraceptive use, age, family history of breast and ovarian cancer, personal history of breast cancer, history of gynecological surgery, and ethnicity. For the purpose of this analysis, we include all women who have complete environmental information and who do not have a bilateral oophorectomy. These eliminations account for about 150 women, leaving 882 cases and 2257 controls available for study. It is important to note that 50 cases and 1510 controls do not have genotype information, and thus could not be utilized by other methods of analysis.

6.2 Analysis

Previous studies have indicated that both parity and oral contraceptive use are important risk factors, with an increase in either being associated with a decreased risk of disease. We now wish to determine if either of these factors interact with the BRCA1/2 mutations. Unfortunately, we believe that many of the other covariates either may not be independent of genotype, may effect the probability of disease, or both. Thus, we will consider these variables as defining a variety of strata. This process defines a large number of strata, and thus we will model the probability of a

mutation as

$$\text{logit}\{\text{pr}(G = 1|\mathbf{S})\} = \theta_{\mathbf{0}} + \theta_{\mathbf{S}}\mathbf{S}.$$

We will also include a linear term for each variable in $m(\cdot)$, as well as the interaction terms between parity and mutation and oral contraceptive use and mutation. Additionally, we will fit the same model to the data assuming marginal independence between the genotype and the environment, but not including any of the women who have genotypes missing.

6.3 Results

The results of our analysis for the parameters of interest, parity, oral contraceptive use, BRCA1/2 mutation, and the interactions between the mutation and each of the other factors, are presented in Table 5. As anticipated, the mutation drastically increases the probability of disease. Also, as hypothesized, increased parity and increased oral contraceptive use decrease the probability of disease. However, oral contraceptive use does not appear to decrease the probability of disease in individuals carrying a BRCA1/2 mutation.

Table 5. Parameter estimates and approximate standard errors for the parameters of interest for the Israeli ovarian cancer study. The first two columns represent the results of an analysis including all individuals who are available for study, regardless of whether or not they have genotype measured. The third and forth columns contain the estimates of the same parameters using only the individuals with known genotype.

Parameter	Estimate	Standard Error	Estimate	Standard Error
β_{mut}	3.01	0.332	2.99	0.468
β_{par}	-0.126	0.049	-0.188	0.070
β_{oc}	-0.185	0.069	-0.139	0.084
$\beta_{\text{mut,par}}$	0.025	0.110	0.023	0.156
$\beta_{\text{mut,oc}}$	0.208	0.098	0.187	0.138

We compare the results of this analysis to an analysis of the data using the assumption of marginal independence of gene and environment, but including only individuals having complete information measured. The results are quite striking; the standard errors for all parameters of interest are significantly smaller under the analysis which incorporates information from individuals with missing genetic information. Additionally, at $\alpha = 0.05$, the effects of both oral contraceptive use and the interaction between oral contraceptive use and the mutation are significant in the case where all of the data are used, but not when only complete individuals are included.

CHAPTER VII

CONCLUSIONS AND FURTHER RESEARCH

In this dissertation, we have developed new methods for both identifying genetic locations in the genome which are associated with a disease of interest and for determining the magnitude of interactions between these genes and environmental factors of interest. These methods provide certain advantages over previous methods for performing similar studies.

The new method for using a joint linkage disequilibrium and linkage analysis to map QTLs permits information from all types of relatives to be used in a joint analysis. This is of particular value in rare human diseases, where finding families that possess a trait of interest is difficult. Furthermore, our method allows all members of these types of families to be included in the study, not just siblings or members of a nuclear family.

A further line of investigation in this area involves the incorporation of information about ascertainment into the analysis. A more extensive likelihood analysis of these types of data which includes ascertainment could possibly provide more complete and accurate information about the locations of QTLs of interest to a particular disease.

The methods developed here for the estimation of gene-environment interactions in case-control studies for logistic regression have many desirable properties, including that they are more flexible, have smaller standard errors, and have known asymptotic properties. These methods make using case-control studies to assess gene-environment interactions feasible in more situations than ever before. Additionally, the assumption of marginal independence between the gene and the environment is

reasonable for many situations.

There are a number of interesting extensions to this work as well. First, there is the question of the effect of measurement error in the environmental variables. A method which could account for this type of measurement error could prove very valuable in a variety of situations; for example, it could prove useful in studies which involve a dietary component and the development of cancer in humans.

A final avenue of possible research involves the extension of these types of methods to study designs other than the case-control study. For example, case-only study designs are popular for this type of research. Analogous methods for other types of studies would increase the flexibility of studies for gene-environment interactions where the gene and environment are marginally independent.

REFERENCES

Abecasis, G. R., Cookson, W. O. C., and Cardon, L. R. (2000), "Pedigree Tests of Linkage Disequilibrium," *Euro J Hum Genet* 8, 545–551.

Abecasis, G. R., Cookson, W. O. C., and Cardon, L. R. (2001), "The Power to Detect Linkage Disequilibrium with Quantitative Traits in Selected Samples," *American Journal of Human Genetics* 68, 1463–1474.

Allison, D. B., Bonnie, T., St. Jean, P., Elston, R. C., Infante, M. C. and Schork, N. J. (1998), "Multiple Phenotype Modeling in Gene-Mapping Studies of Quantitative Traits: Power Advantages," *American Journal of Human Genetics* 63, 1190–1201.

Almasy, L. and Blangero, J. (1998), "Multipoint Quantitative Trait Linkage Analysis in General Pedigrees," *American Journal of Human Genetics* 62, 1198–1211.

Almasy, L., Williams, J. T., Dyer, T. D., and Blangero, J. (1999), "Quantitative Trait Locus Detection Using Combined Linkage/Disequilibrium Analysis," *Genetic Epidemiology* 17, S31–S36.

Amos, C. I. (1994), "Robust Variance-Components Approach for Assessing Linkage in Pedigrees," *American Journal of Human Genetics* 54, 534–543.

Amos, C. I., and Elston, R. C. (1989), "Robust Methods for the Detection of Genetic Linkage for Quantitative Data from Pedigrees," *Genetic Epidemiology* 6, 349–360.

Amos, C. I., Elston, R. C., Wilson, A. F., and Bailey-Wilson, J. E. (1989), "A More Powerful Robust Sib-Pair Test of Linkage for Quantitative Traits," *Genetic Epidemiology* 6, 435–449.

Broman, K. W., Murray, J. C., Sheffied, V. C., White, R. L. and Weber, J. L. (1998),

"Comprehensive Human Genetic Map: Individual and Sex-Specific Variation in Recombination," *American Journal of Human Genetics* 63, 861–869.

Cardon, L. R. (2000), "A Sib-Pair Regression Model of Linkage Disequilibrium for Quantitative Traits," *Human Heredity* 50, 350–358.

Chatterjee, N. and Carroll, R. J., " Semiparametric Maximum Likelihood Estimation in Case-Control Studies of Gene-Environment Interactions," *Biometrika* (to appear).

Clark, A. G. (1990), "Inference of Haplotypes from PCR-amplified Samples of Diploid Populations," *Molecular Biology and Evolution* 7, 111–122.

Excoffier, L. and Slatkin, M. (1995). " Maximum-likelihood Estimation of Molecular Haplotype Frequencies in a Diploid Population," *Molecular Biology and Evolution* 12, 921–927.

Falconer, D. S. and Mackay, T. F. C. (1996), *Introduction to Quantitative Genetics*, London: Longman.

Fallin, D. and Schork, N. J. (2000), "Accuracy of Haplotype Frequency Estimation for Biallelic Loci, via the Expectation-Maximization Algorithm for Unphased Diploid Genotype Data," *American Journal of Human Genetics* 67, 947–959.

Fan, R. and Jung, J. (2003), "High Resolution Joint Linkage Disequilibrium and Linkage Mapping of Quantitative Trait Loci Based on Sibship Data," *Human Heredity* 56, 166–187.

Fan, R. and Xiong, M. (2002), "High Resolution Mapping of Quantitative Trait Loci by Linkage Disequilibrium Analysis," *European Journal of Human Genetics* 10, 607–615.

Fan, R. and Xiong, M. (2003), "Combined High Resolution Linkage and Association Mapping of Quantitative Trait Loci," *European Journal of Human Genetics* 11, 125–137.

Fulker, D. W., Cherny, S. S. and Cardon, L. R. (1995), " Multiple Interval Mapping of Quantitative Trait Loci, Using Sib-Pairs," *American Journal of Human Genetics* 56, 1224–1233.

Fulker, D. W., Cherny, S. S., Sham, P. C., and Hewitt, J. K. (1999), "Combined Linkage and Association Sib-Pair Analysis for Quantitative Traits," *American Journal of Human Genetics* 64, 259–267.

Goldgar, D. E. and Oniki, R. S. (1992), "Comparison of a Multipoint Identity-by-Descent Method with Parametric Multipoint Linkage Analysis for Mapping Quantitative Traits," *American Journal of Human Genetics* 50, 598–606.

Göring, H. H. H. and Terwilliger, J. D. (2000), "Linkage Analysis in the Presence of Error IV: Joint Pseudomarker Analysis of Linkage and/or Linkage Disequilibrium on a Mixture of Pedigrees and Singletons When the Mode of Inheritance Cannot be Accurately Specified," *American Journal of Human Genetics* 66, 1310–1327.

Graybill, F. A. (1976), *Theory and Application of the Linear Model*, Pacific Grove, California.

Kong, A., Gudbjartsson, D. F., Sainz, J., Jonsdottir, G. M., Gudjonsson, S. A. et al. (2002), "A High Resolution Recombination Map of the Human Genome," *Nature Genetics* 31, 241–247.

Martin, E. R., Monks, S. A., Warren, L. L., and Kaplan, N. L. (2000), "A Test for Linkage and Association in General Pedigrees: The Pedigree Disequilibrium Test," *American Journal of Human Genetics* 67, 146–154.

Niu, T., Qin, Z. S., Xu, X. and Liu, J. S. (2002), "Bayesian Haplotype Inference for Multiple Linked Single- Nucleotide Polymorphisms," *American Journal of Human Genetics* 70, 157–169.

Pratt, S. C., Daly, M., and Kruglyak, L. (2000), "Exact Multipoint Quantitative-Trait Linkage Analysis in Pedigrees by Variance Components," *American Journal of*

Human Genetics 66, 1153–1157.

Prentice, R. L. and Pyke, R. (1979), "Logistic Disease Incidence Models and Case-control Studies," *Biometrika* 66, 403–412.

Roeder, K., Carroll, R. J. and Lindsay, B. G. (1996), " A Semiparametric Mixture Approach to Case-control Studies with Errors in Covariables," *Journal of the American Statistical Association* 91, 722–732.

Satten, G. A., Flanders, W. D. and Yang, Q. (2001), " Accounting for Unmeasured Population Substructure in Case-Control Studies of Genetic Association Using a Novel Latent- Class Model," *American Journal of Human Genetics* 68, 466–477.

Schaid, D. J., Rowland, C. M., Tines, D. E., Jacobson, R. M. and Poland, G. A. (2002), "Score Tests for Association between Traits and Haplotypes When Linkage Phase Is Ambiguous," *American Journal of Human Genetics* 70, 425–434.

Sham, P. C., Cherny, S. S., Purcell, S. and Hewitt, J. K. (2000), "Power of Linkage versus Association Analysis of Quantitative Traits, by Use of Variance-Components Models, for Sibship Data," *American Journal of Human Genetics* 66, 1616–1630.

Thomas, D. C., Morrison, J. L. and Clayton, D. G. (2001), "Bayes Estimates of Haplotype Effects," *Genetic Epidemiology* 21, S712–S717.

Umbach, D. M. and Weinberg, C. R. (1997), "Designing and Analysing Case-control Studies to Exploit Independence of Genotype and Exposure," *Statistics in Medicine* 16, 1731–1743.

Wallenstein, S., Hodge, S. E. and Weston, A. (1998), "Logistic Regression Model for Analyzing Extended Haplotype Data," *Genetic Epidemiology* 15, 173–183.

Wong, M. Y., Day, N. E., Luan, J. A. and Wareham, N. J. (2004), "Estimation of Magnitude in Gene-Environment Interactions in the Presence of Measurement Error," *Statistics in Medicine* 23, 987–998.

Yang, Q. and Khoury, M. J. (1997), "Evolving Methods in Genetic Epidemiology.

III. Gene-Environment Interaction in Epidemiologic Research," *Epidemiologic Reviews* 19, 33–43.

Zhao, L. P., Li, S. S. and Khalid, N. (2003), "A Method for the Assessment of Disease Associations with Single-Nucleotide Polymorphism Haplotypes and Environmental Variables in Case- Control Studies," *American Journal of Human Genetics* 72, 1231–1250.

APPENDIX A

PROOFS FOR CHAPTER II

A.1 Determining Expected Covariance Matrices

For a relative pair $(1, 2)$, Table 6 gives the conditional probability $P(G_1, G_2|C)$ given their allele IBD sharing status. Here G_i is genotype of relative i, and C is one event of $(IBD = k), k = 0, 1, 2$. For example, $P(AA, AA|IBD = 0) = P_A^4$, $P(AA, AA|IBD = 1) = P_A^3$ and $P(AA, AA|IBD = 2) = P_A^2$. Let us denote the additive and dominance functions

$$x_{Ai} = \begin{cases} 2P_a & \text{if} A_i = AA \\ P_a - P_A & \text{if} A_i = Aa \\ -2P_A & \text{if} A_i = aa \end{cases}, \qquad z_{Ai} = \begin{cases} -P_a^2 & \text{if} A_i = AA \\ P_a P_A & \text{if} A_i = Aa \\ -P_A^2 & \text{if} A_i = aa \end{cases}.$$

Define the functions for locus B similarly.

Then, utilizing the conditional probabilities of Table 6, the conditional expectation of the product of the additive and dominance functions can be calculated and the results are listed in Table 7. For instance, $\mathrm{E}(x_{A1}x_{A2}|IBD = 2) = 4P_a^2P_A^2 + (P_a - P_A)^2 \cdot 2P_aP_A + 4P_A^2P_a^2 = 2P_aP_A$. Based on Table 7, the expectation of the product of these functions can be calculated for various types of relative pairs.

For example, if individuals 1 and 2 are the same person, $P(IBD = 2) = 1$ and then

$$\mathrm{E}\Big[(1, x_{A1}, x_{B1}, z_{A1}, z_{B1})^\tau (1, x_{A2}, x_{B2}, z_{A2}, z_{B2})\Big] = \begin{pmatrix} 1 & O & O \\ O & V_A & O \\ O & O & V_D \end{pmatrix},$$

where O are 0 matrices. If individuals 1 and 2 are heterozygous sibs, $P(IBD = 0) =$

Table 6. Conditional probability $P(G_1, G_2|C)$ of a relative pair $(1, 2)$ given their allele IBD sharing status. Here G_i is genotype of relative i, and C is one event of $(IBD = k), k = 0, 1, 2$.

Conditional Probability	**allele IBD sharing status C**		
	IBD=0	**IBD=1**	**IBD=2**
$P(AA, AA\|C)$	P_A^4	P_A^3	P_A^2
$P(AA, Aa\|C)$	$2P_aP_A^3$	$P_aP_A^2$	0
$P(AA, aa\|C)$	$P_a^2P_A^2$	0	0
$P(Aa, AA\|C)$	$2P_aP_A^3$	$P_aP_A^2$	0
$P(Aa, Aa\|C)$	$4P_a^2P_A^2$	$P_aP_A^2 + P_a^2P_A$	$2P_aP_A$
$P(Aa, aa\|C)$	$2P_a^3P_A$	$P_a^2P_A$	0
$P(aa, AA\|C)$	$P_a^2P_A^2$	0	0
$P(aa, Aa\|C)$	$2P_a^3P_A$	$P_a^2P_A$	0
$P(aa, aa\|C)$	P_a^4	P_a^3	P_a^2
$P(AA, BB\|C)$	$P_A^2P_B^2$	$P_AP_BP(AB)$	$P(AB)^2$
$P(AA, Bb\|C)$	$2P_A^2P_bP_B$	$P_AP_bP(AB) + P_AP_BP(Ab)$	$2P(AB)P(Ab)$
$P(AA, bb\|C)$	$P_A^2P_b^2$	$P_AP_bP(Ab)$	$P(Ab)^2$
$P(Aa, BB\|C)$	$2P_aP_AP_B^2$	$P_AP_BP(aB) + P_aP_BP(AB)$	$2P(AB)P(aB)$
$P(Aa, Bb\|C)$	$4P_aP_AP_bP_B$	$P_AP_BP(ab) + P_AP_bP(aB)$ $+P_aP_BP(Ab) + P_aP_bP(AB)$	$2P(AB)P(ab)$ $+2P(Ab)P(aB)$
$P(Aa, bb\|C)$	$2P_aP_AP_b^2$	$P_AP_bP(ab) + P_aP_bP(Ab)$	$2P(Ab)P(ab)$
$P(aa, BB\|C)$	$P_a^2P_B^2$	$P_aP_BP(aB)$	$P(aB)^2$
$P(aa, Bb\|C)$	$2P_a^2P_bP_B$	$P_aP_bP(aB) + P_aP_BP(ab)$	$2P(aB)P(ab)$
$P(aa, bb\|C)$	$P_a^2P_b^2$	$P_aP_bP(ab)$	$P(ab)^2$

$P(IBD = 2) = 1/4, P(IBD = 1) = 1/2$ and then

$$\mathrm{E}\Big[(1, x_{A1}, x_{B1}, z_{A1}, z_{B1})^\tau (1, x_{A2}, x_{B2}, z_{A2}, z_{B2})\Big] = \begin{pmatrix} 1 & O & O \\ O & V_A/2 & O \\ O & O & V_D/4 \end{pmatrix}.$$

If individuals 1 and 2 are a parent-offspring pair, $P(IBD = 0) = P(IBD = 2) =$

Table 7. Conditional expectation of a relative pair $(1,2)$ *given their allele IBD sharing status.*

Conditional Expectation	**allele IBD sharing status** C		
	IBD=0	**IBD=1**	**IBD=2**
$\mathrm{E}(x_{A1}x_{A2}\|C)$	0	P_aP_A	$2P_aP_A$
$\mathrm{E}(x_{A1}z_{A2}\|C)$	0	0	0
$\mathrm{E}(x_{A1}x_{B2}\|C)$	0	D_{AB}	$2D_{AB}$
$\mathrm{E}(x_{A1}z_{B2}\|C)$	0	0	0
$\mathrm{E}(z_{A1}x_{A2}\|C)$	0	0	0
$\mathrm{E}(z_{A1}z_{A2}\|C)$	0	0	$P_a^2P_A^2$
$\mathrm{E}(z_{A1}x_{B2}\|C)$	0	0	0
$\mathrm{E}(z_{A1}z_{B2}\|C)$	0	0	D_{AB}^2

$0, P(IBD=1)=1$ and then

$$\mathrm{E}\Big[(1,x_{A1},x_{B1},z_{A1},z_{B1})^\tau(1,x_{A2},x_{B2},z_{A2},z_{B2})\Big]=\begin{pmatrix}1 & O & O\\ O & V_A/2 & O\\ O & O & O\end{pmatrix}.$$

If individuals 1 and 2 are a grand-parent/grand-child pair (or uncle/niece or aunt/ nephew pair), $P(IBD=0)=1/2$, $P(IBD=2)=0$, $P(IBD=1)=1/2$ and then

$$\mathrm{E}\Big[(1,x_{A1},x_{B1},z_{A1},z_{B1})^\tau(1,x_{A2},x_{B2},z_{A2},z_{B2})\Big]=\begin{pmatrix}1 & O & O\\ O & V_A/4 & O\\ O & O & O\end{pmatrix}.$$

If individuals 1 and 2 are first cousins, $P(IBD=0)=3/4, P(IBD=2)=0, P(IBD=1)=1/4$ and then

$$\mathrm{E}\Big[(1,x_{A1},x_{B1},z_{A1},z_{B1})^\tau(1,x_{A2},x_{B2},z_{A2},z_{B2})\Big]=\begin{pmatrix}1 & O & O\\ O & V_A/8 & O\\ O & O & O\end{pmatrix}.$$

A.2 Estimating Non-Centrality Parameters

Based on the equations found in Section A.1, we can approximate the non-centrality parameter of a test statistic, F, as follows. Denote $\Sigma_i^{-1} = \frac{1}{\sigma^2}(\gamma_{kl})_{n\times n}$, $n = 11$ for graph A and $n = 18$ for graph B of Figure 1, respectively. When I is sufficiently large, the strong law of large numbers implies that

$$\frac{1}{I}\sum_{i=1}^{I} X_i^{\tau}{\Sigma_i}^{-1}X_i \approx \frac{1}{\sigma^2}\begin{pmatrix} \sum_k\sum_l \gamma_{kl} & O & O \\ O & b_1V_A & O \\ O & O & b_2V_D \end{pmatrix}.$$

Here, b_1 and b_2 are constants as follows for pedigrees in graph A of Figure 1

$$\begin{aligned} b_1 &= \sum_{k=1}^{11}\gamma_{kk} + [\gamma_{15} + (\gamma_{17} + \cdots + \gamma_{1,11})/2] + [\gamma_{25} + (\gamma_{27} + \cdots + \gamma_{2,11})/2] \\ &\quad + [\gamma_{36} + (\gamma_{37} + \cdots + \gamma_{3,11})/2] + [\gamma_{46} + (\gamma_{47} + \cdots + \gamma_{4,11})/2] \\ &\quad + (\gamma_{57} + \cdots + \gamma_{5,11}) + (\gamma_{67} + \cdots + \gamma_{6,11}) + \sum_{k=7}^{11}\sum_{l=k+1}^{11}\gamma_{kl}, \\ b_2 &= \sum_{k=1}^{11}\gamma_{kk} + \sum_{k=7}^{11}\sum_{l=k+1}^{11}\gamma_{kl}/2. \end{aligned}$$

For pedigrees in graph B of Figure 1, constants b_1 and b_2 are given by

$$\begin{aligned}
b_1 &= \sum_{k=1}^{18}\gamma_{kk}+\gamma_{1,12}+[\gamma_{2,12}+(\gamma_{2,13}+\cdots+\gamma_{2,16})/2]+[\gamma_{3,12}+\cdots+\gamma_{3,16}]/2 \\
&\quad +[\gamma_{4,12}+\cdots+\gamma_{4,16}]/2+[\gamma_{5,12}/2+(\gamma_{5,13}+\cdots+\gamma_{5,16})] \\
&\quad +[(\gamma_{6,13}+\cdots+\gamma_{6,16})+(\gamma_{6,17}+\gamma_{6,18})/2]+[\gamma_{7,13}+\cdots+\gamma_{7,18}]/2 \\
&\quad +[(\gamma_{8,13}+\cdots+\gamma_{8,16})/2+(\gamma_{8,17}+\gamma_{8,18})]+(\gamma_{9,17}+\gamma_{9,18}) \\
&\quad +(\gamma_{10,17}+\gamma_{10,18})/2+(\gamma_{11,17}+\gamma_{11,18})/2+(\gamma_{12,13}+\cdots+\gamma_{12,16})/4 \\
&\quad +(\gamma_{13,14}+\gamma_{13,15}+\gamma_{13,16})+(\gamma_{14,15}+\gamma_{14,16})+\gamma_{15,16} \\
&\quad +[\gamma_{13,17}+\cdots+\gamma_{16,17}]/4+[\gamma_{13,18}+\cdots+\gamma_{16,18}]/4+\gamma_{17,18}, \\
b_2 &= \sum_{k=1}^{18}\gamma_{kk}+[\gamma_{13,14}+\gamma_{13,15}+\gamma_{13,16}+\gamma_{14,15}+\gamma_{14,16}+\gamma_{15,16}]/2+\gamma_{17,18}/2.
\end{aligned}$$

Therefore, the non-centrality parameter can be approximated by

$$\begin{aligned}
\lambda &= (H\mu)^\tau\Big[H[\sum_{i=1}^{I}X_i^\tau{\Sigma_i}^{-1}X_i]^{-1}H^\tau\Big]^{-1}(H\mu) \\
&\approx \frac{I}{\sigma^2}(H\mu)^\tau\left[H\begin{pmatrix} 1/\sum_k\sum_l\gamma_{kl} & O & O \\ O & V_A^{-1}/b_1 & O \\ O & O & V_D^{-1}/b_2 \end{pmatrix}H^\tau\right]^{-1}(H\mu).
\end{aligned}$$

APPENDIX B

PROOFS FOR CHAPTER III

B.1 Calculating the Profile Likelihood

Roeder et al. (1996) show that the parameters of interest in the logistic regression are identified. However, the dimension of the environmental distribution grows with the sample size. Additionally, it is not of primary interest to characterize the distribution of the environmental parameters. Thus, the profile likelihood provides a means of determining the estimates for the parameters of interest without needing to explicitly maximize the environmental parameters. First, for a fixed value of $\gamma = (\beta_0, \beta_1, \theta)$ the likelihood function for ζ has the form

$$\ell(\zeta|\gamma) \quad \propto \quad \sum_{i=1}^{n} \log\{\zeta(X_i)\} - \sum_{i=1}^{n} \log\left\{ \sum_{j,k} \text{pr}(D = D_i | X = x_k, G = g_j) h_j(\theta) \zeta_k \right\},$$

as the other terms are constants. Then, taking derivatives with respect to each ζ_m and solving for the MLEs of the ζ's,

$$\ell_{\zeta_m}(\zeta|\gamma) \quad = \quad \frac{\sum_i I(X_i = x_m)}{\zeta_m} - \sum_{i=1}^{n} \frac{\sum_j \text{pr}(D = D_i | X = x_m, G = g_j) h_j(\theta)}{\sum_{j,k} \text{pr}(D = D_i | X = x_k, G = g_j) h_j(\theta) \zeta_k}.$$

At the maximum likelihood estimate,

$$\zeta_m \quad = \quad \sum_i I(X_i = x_m) / \sum_{i=1}^{n} \frac{\sum_j \text{pr}(D = D_i | X = x_m, G = g_j) h_j(\theta)}{\sum_{j,k} \text{pr}(D = D_i | X = x_k, G = g_j) h_j(\theta) \zeta_k}.$$

However, notice that $\text{pr}(D = d) = \sum_{j,k} \text{pr}(D = d | X = x_k, G = g_j) h_j(\theta) \zeta_k$, and define $\mu_d = n_d / \{n \text{pr}(D = d)\}$. This implies that $\text{pr}(D = d) = n_d / (n \mu_d)$ and

$$\zeta_m \quad = \quad \frac{\sum_i I(X_i = x_m)}{n \sum_{d,j} \text{pr}(D = d | X = x_\ell, G = g_j) \mu_d h_j(\theta)}.$$

Then, $\zeta_m \sum_{d,j} \text{pr}(D = d|X = x_m, G = g_j)\mu_d h_j(\theta) = n^{-1} \sum_i I(X_i = x_m)$, which implies that $\sum_{d,j,m} \zeta_m \text{pr}(D = d|X = x_m, G = g_j)\mu_d h_j(\theta) = 1$. Now,

$$\mu_0 \sum_{j,m} \zeta_m \text{pr}(D = 0|x_m, g_j)h_j(\theta) + \mu_1 \sum_{j,m} \zeta_m \left\{1 - \text{pr}(D = 0|x_m, g_j)\right\} h_j(\theta) \;=\; 1,$$

$$(\mu_0 - \mu_1) \sum_{j,m} \zeta_m \text{pr}(D = 0|X = x_m, G = g_j)h_j(\theta) + \mu_1 \sum_{j,m} \zeta_m h_j(\theta) \;=\; 1,$$

and $(\mu_0 - \mu_1)\text{pr}(D = 0) + \mu_1 \sum_m \zeta_m = 1$, which implies that $\sum_m \zeta_m = 1$, as is desired under this model. Thus, the profile likelihood function has the form

$$\begin{aligned}
\ell\{\gamma, \zeta(\gamma)\} \;=\; & \sum_{i=1}^{n} \Bigg[\log\{\pi(\Delta_i|D_i, X_i, G_i^*)\} + \log \left\{ \sum_{g_j \in \mathcal{G}_i} \text{pr}(D_i|X_i, g_j)h(g_j, \theta) \right\} \\
& + \log \left\{ \sum_l I(X_l = X_i) \right\} - \log \left\{ n \sum_{d,j} \text{pr}(d|X_i, g_j)\mu_d h(g_j, \theta) \right\} \\
& + \log(n_{D_i}) - \log\{n\mu(D_i)\} \Bigg] \\
\propto & \sum_{i=1}^{n} \left[\log \left\{ \sum_{g_j \in \mathcal{G}_i} \text{pr}(D_i|X_i, g_j)h(g_j, \theta) \right\} + \log\{\mu(D_i)\} \right. \\
& \left. - \log \left\{ \sum_{d,j} \text{pr}(d|X_i, g_j)\mu_d h(g_j, \theta) \right\} \right].
\end{aligned}$$

Now, define $\kappa = \beta_0 + \log(\mu_1/\mu_0)$ or equivalently $\mu_1 = \mu_0 \exp(\kappa - \beta_0)$ or $\log(\mu_1) = \log(\mu_0) + (\kappa - \beta_0)$. Then,

$$\begin{aligned}
\ell\{\gamma, \zeta(\gamma)\} \;=\; & \sum_{i=1}^{n} \left(\log \left[\sum_{g_j \in \mathcal{G}_i} \text{pr}(D_i|X_i, g_j)h(g_j, \theta) \exp\{d(\kappa - \beta_0)\} \right] \right. \\
& \left. - \log \left[\sum_{d,j} \text{pr}(d|X_i, g_j)h(g_j, \theta) \exp\{d(\kappa - \beta_0)\} \right] \right).
\end{aligned}$$

Now, define $\Omega = (\gamma, \kappa)$, $H(x) = \{1 + \exp(-x)\}^{-1}$, and

$$S(D, X, G, \Omega) \;=\; h(G, \theta) \exp[D\{\kappa + m(X, G, \beta_1)\}][1 - H\{\beta_0 + m(X, G, \beta_1)\}].$$

Notice that

$$\begin{aligned} S(1,X,G,\Omega) &= h(G,\theta)\exp(\kappa-\beta_0)\exp\{beta_0+m(X,G,\beta_1)\} \\ &\quad \times[1-H\{\beta_0+m(X,G,\beta_1)\}] \\ &= h(G,\theta)\exp(\kappa-\beta_0)\mathrm{pr}(D=1|X,G) \end{aligned}$$

and $S(0,X,G,\Omega) = h(G,\theta)[1-H\{\beta_0+m(X,G,\beta_1)\}] = h(G,\theta)\mathrm{pr}(D=0|X,G)$. Thus, we can write the profile likelihood as

$$\ell\{\gamma,\zeta(\gamma)\} = \sum_{i=1}^{n}\left[\log\left\{\sum_{g_j\in\mathcal{G}_i} S(D_i,X_i,g_j,\Omega)\right\} - \log\left\{\sum_{d,j} S(d,X_i,g_j,\Omega)\right\}\right],$$

as was to be shown.

B.2 Asymptotic Distribution of the Estimates

B.2.1 The Score

To understand the distribution of the estimates derived from the profile likelihood function, we need to first construct the score. First, define $\ell_\Omega(\Omega) = \partial/\partial\Omega\{\ell(\Omega)\}$ and $S_\Omega(D,X,G,\Omega) = \partial/\partial\Omega\{S(D,X,G,\Omega)\}$. Now, the score of the profile likelihood has the form

$$\begin{aligned} \ell_\Omega(\Omega) &= \sum_{i=1}^{n}\frac{\partial}{\partial\Omega}\left[\log\left\{\sum_{g_j\in\mathcal{G}_i} S(D_i,X_i,g_j,\Omega)\right\} - \log\left\{\sum_{d,j} S(d,X_i,g_j,\Omega)\right\}\right] \\ &= \sum_{i=1}^{n}\frac{\sum_{g_j\in\mathcal{G}_i} S_\Omega(D_i,X_i,g_j,\Omega)}{\sum_{g_j\in\mathcal{G}_i} S(D_i,X_i,g_j,\Omega)} - \sum_{i=1}^{n}\frac{\sum_{d,j} S_\Omega(d,X_i,g_j,\Omega)}{\sum_{d,j} S(d,X_i,g_j,\Omega)}. \end{aligned}$$

Lemma A.1: The score of the profile likelihood is unbiased, and thus can be considered as a set of unbiased estimating equations.

The following lemma is useful in studying the distributional properties of the estimates obtained from the profile likelihood.

Lemma A.2: For any function $R(\Delta, D, X, G^*)$,

$$\mathrm{E}\left\{\sum_{i=1}^{n} R(\Delta_i, D_i, X_i, G_i^*)\right\} = \frac{n_0}{\mathrm{pr}(D=0)} \int_x f_X(x) \sum_{d,j,l} R(\Delta_l, d, x, g_j^*) S(d, x, g_j, \Omega) \times \pi(\Delta_l | d, x, g_j^*) dx.$$

Corollary: For any function $R(D, X, G)$,

$$\mathrm{E}\left\{\sum_{i=1}^{n} R(D_i, X_i, G_i)\right\} = \frac{n_0}{\mathrm{pr}(D=0)} \int_x f_X(x) \sum_{d,j} R(d, x, g_j) S(d, x, g_j, \Omega) dx.$$

Proof of Lemma A.2: Notice,

$$\begin{aligned}
\mathrm{E}\left\{\sum_{i=1}^{n} R(\Delta_i, D_i, X_i, G_i^*)\right\} &= \frac{n_1}{\mathrm{pr}(D=1)} \int_x \sum_{j,\ell} R(\Delta_\ell, 1, x, g_j^*) \\
&\quad \times \pi(\Delta_l | 1, x, g_j^*) f(1 | x, g_j^*) h(g_j^*, \theta) f_X(x) dx \\
&\quad + \frac{n_0}{\mathrm{pr}(D=0)} \int_x \sum_{j,\ell} R(\Delta_\ell, 0, x, g_j^*) \pi(\Delta_l | 0, x, g_j^*) f(0 | x, g_j^*) h(g_j^*, \theta) f_X(x) dx \\
&= \frac{n_0}{\mathrm{pr}(D=0)} \int_x \sum_{j,\ell} R(\Delta_\ell, 1, x, g_j^*) S(1, x, g_j, \Omega) \pi(\Delta_l | 1, x, g_j^*) f_X(x) dx \\
&\quad + \frac{n_0}{\mathrm{pr}(D=0)} \int_x \sum_{j,\ell} R(\Delta_\ell, 0, x, g_j^*) S(0, x, g_j, \Omega) \pi(\Delta_l | 0, x, g_j^*) f_X(x) dx \\
&= \frac{n_0}{\mathrm{pr}(D=0)} \int_x f_X(x) \sum_{d,j,l} R(\Delta_l, d, x, g_j^*) S(d, x, g_j, \Omega) \pi(\Delta_l | d, x, g_j^*) dx.
\end{aligned}$$

The corollary follows immediately from the fact that $\sum_\ell \pi(\Delta_\ell | d, X, g_j) = 1$.

Proof of Lemma A.1: Notice that

$$\begin{aligned}
\ell_\Omega(\Omega) &= \frac{n_0}{\mathrm{pr}(D=0)} \int_x f_X(x) \sum_{d,j,\ell} \frac{\sum_{g_j \in \mathcal{G}_\ell} S_\Omega(d, x, g_j, \Omega)}{\sum_{g_j \in \mathcal{G}_l} S(d, x, g_j, \Omega)} S(d, x, g_j, \Omega) \pi(\Delta_\ell | d, x, g_j^*) dx \\
&\quad - \frac{n_0}{\mathrm{pr}(D=0)} \int_x f_X(x) \sum_{d,j} \frac{\sum_{d,j} S_\Omega(d, x, g_j, \Omega)}{\sum_{d,j} S(d, x, g_j, \Omega)} S(d, x, g_j, \Omega) dx \\
&= \frac{n_0}{\mathrm{pr}(D=0)} \int_x f_X(x) \sum_{d,j} S_\Omega(d, x, g_j, \Omega) dx \\
&\quad - \frac{n_0}{\mathrm{pr}(D=0)} \int_x f_X(x) \sum_{d,j} S_\Omega(d, x, g_j, \Omega) dx.
\end{aligned}$$

Thus, $\mathrm{E}\{\ell_\Omega(\Omega)\} = 0$ by an application of Lemma A.2.

B.2.2 Asymptotic Distributional Results

Lemma A.3: For the case–control study design, with fixed n_0/n, measurable functions $R(\Delta, D, X, G^*)$ satisfy

$$\frac{1}{n}\sum_{i=1}^{n} R(\Delta_i, D_i, X_i, G_i^*) \longrightarrow_P \mu_0 \int_x f_X(x) \sum_{d,j,l} R(\Delta_l, d, x, g_j^*) S(d, x, g_j, \Omega) \times \pi(\Delta_l | d, x, g_j^*) dx,$$

assuming that the integral exists.

Proof of Lemma A.3: Notice

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} R(\Delta_i, D_i, X_i, G_i^*) &= \frac{1}{n_0}\frac{n_0}{n}\sum_{i:d_i=0} R(\Delta_i, D_i = 0, X_i, G_i^*) \\ &\quad + \frac{1}{n_1}\frac{n_1}{n}\sum_{i:d_i=1} R(\Delta_i, D_i = 1, X_i, G_i^*).\end{aligned}$$

Now, the cases, $d_i = 1$, are iid from the distribution of cases, and the controls, $d_i = 0$, are iid from their distribution. Also,

$$n_d^{-1}\sum_{i:d_i=d} R(\Delta_i, D_i, X_i, G_i^*) \longrightarrow_P \mathrm{E}\left\{R(\Delta, D, X, G^*) | D = d\right\},$$

by the Weak Law of Large Numbers. Thus

$$\begin{aligned}n^{-1}\sum_{i=1}^{n} R(\Delta_i, D_i, X_i, G_i^*) \longrightarrow_P\ & n_0 n^{-1}\mathrm{E}\{R(\Delta, D, X, G^*) | D = 0\} \\ & + n_1 n^{-1}\mathrm{E}\{R(\Delta, D, X, G^*) | D = 1\}.\end{aligned}$$

Also, from Lemma A.2,

$$\begin{aligned}\mathrm{E}\{R(\Delta, D, X, G^*) | D = d\} &= n\mu_0 n_d^{-1} \int_x f_X(x) \sum_{j,l} R(\Delta_l, d, x, g_j^*) S(d, x, g_j, \Omega) \\ &\quad \times \pi(\Delta_l | d, x, g_j^*) dx.\end{aligned}$$

Thus,

$$n^{-1}\sum_{i=1}^{n} R(\Delta_i, D_i, X_i, G_i^*) \longrightarrow_P \mu_0 \int_x f_X(x) \sum_{d,j,l} R(\Delta_l, d, x, g_j^*) S(d, x, g_j, \Omega) \times \pi(\Delta_l | d, x, g_j^*) dx.$$

The Matrix of Second Partials: Define $S_{\Omega\Omega^T}(D,X,G,\Omega)$ to be the matrix of second derivatives of $S(D,X,G,\Omega)$ with respect to Ω. Define $\ell_{\Omega\Omega^T}(\Omega)$ similarly. Then, note that

$$\begin{aligned}\ell_{\Omega\Omega^T}(\Omega) &= \sum_{i=1}^{n}\left[\frac{\sum_{g_j\in\mathcal{G}_i} S_{\Omega\Omega^T}(D_i,X_i,g_j,\Omega)}{\sum_{g_j\in\mathcal{G}_i} S(D_i,X_i,g_j,\Omega)} - \frac{\sum_{d,j} S_{\Omega\Omega^T}(d,X_i,g_j,\Omega)}{\sum_{dj} S(d,X_i,g_j,\Omega)}\right.\\ &\quad -\frac{\sum_{g_j\in\mathcal{G}_i} S_{\Omega}(D_i,X_i,g_j,\Omega)\{\sum_{g_j\in\mathcal{G}_i} S_{\Omega}(D_i,X_i,g_j,\Omega)\}^T}{\{\sum_{g_j\in\mathcal{G}_i} S(D_i,X_i,g_j,\Omega)\}^2}\\ &\quad \left.+\frac{\sum_{d,j} S_{\Omega}(d,X_i,g_j,\Omega)\{\sum_{d,j} S_{\Omega}(d,X_i,g_j,\Omega)\}^T}{\{\sum_{dj} S(d,X_i,g_j,\Omega)\}^2}\right].\end{aligned}$$

Consider the first and second terms denoted by S_{n1} and S_{n2}, respectively. Then,

$$\begin{aligned}\mathrm{E}(S_{n1}) &= \frac{n_0}{\mathrm{pr}(D=0)}\int_x f_X(x)\sum_{d,j,\ell}\frac{\sum_{g_j\in\mathcal{G}_\ell} S_{\Omega\Omega^T}(d,x,g_j,\Omega)}{\sum_{g_j\in\mathcal{G}_\ell} S(d,x,g_j,\Omega)}S(d,x,g_j,\Omega)\\ &\quad\times\pi(\Delta_\ell|d,x,g_j^*)dx\\ &= \frac{n_0}{\mathrm{pr}(D=0)}\int_x f_X(x)\sum_{d,j} S_{\Omega\Omega^T}(d,x,g_j,\Omega)dx\\ \mathrm{E}(S_{n2}) &= \frac{n_0}{\mathrm{pr}(D=0)}\int_x f_X(x)\sum_{d,j}\frac{\sum_{d,j} S_{\Omega\Omega^T}(d,x,g_j,\Omega)}{\sum_{dj} S(d,x,g_j,\Omega)}S(d,x,g_j,\Omega)dx\\ &= \frac{n_0}{\mathrm{pr}(D=0)}\int_x f_X(x)\sum_{d,j} S_{\Omega\Omega^T}(d,x,g_j,\Omega)dx.\end{aligned}$$

Then $\mathrm{E}(S_{n1}-S_{n2})=0$, and

$$\begin{aligned}-\mathrm{E}\left[\ell_{\Omega\Omega^T}(\Omega)\right] &= \mathrm{E}\left[\sum_{i=1}^{n}\frac{\sum_{g_j\in\mathcal{G}_i} S_{\Omega}(D_i,X_i,g_j,\Omega)\{\sum_{g_j\in\mathcal{G}_i} S_{\Omega}(D_i,X_i,g_j,\Omega)\}^T}{\{\sum_{g_j\in\mathcal{G}_i} S(D_i,X_i,g_j,\Omega)\}^2}\right]\\ &\quad -\mathrm{E}\left[\sum_{i=1}^{n}\frac{\sum_{d,j} S_{\Omega}(d,X_i,g_j,\Omega)\{\sum_{d,j} S_{\Omega}(d,X_i,g_j,\Omega)\}^T}{\{\sum_{dj} S(d,X_i,g_j,\Omega)\}^2}\right]\\ &= \mathrm{E}(C_{n1})-\mathrm{E}(C_{n2}),\end{aligned}$$

where

$$\begin{aligned}\mathrm{E}(C_{n1}) &= \frac{n_0}{\mathrm{pr}(D=0)}\int_x f_X(x)\sum_{d,j,\ell}\frac{\sum_{g_j\in\mathcal{G}_\ell} S_{\Omega}(d,x,g_j,\Omega)\{\sum_{g_j\in\mathcal{G}_\ell} S_{\Omega}(d,x,g_j,\Omega)\}^T}{\sum_{g_j\in\mathcal{G}_\ell} S(d,x,g_j,\Omega)}\\ &\quad\times\pi(\Delta|d,x,g_j^*)dx\\ \mathrm{E}(C_{n2}) &= \frac{n_0}{\mathrm{pr}(D=0)}\int_x f_X(x)\frac{\sum_{d,j} S_{\Omega}(d,x,g_j,\Omega)\{\sum_{d,j} S_{\Omega}(d,x,g_j,\Omega)\}^T}{\sum_{dj} S(d,x,g_j,\Omega)}dx.\end{aligned}$$

The Variance of the Score: Recall that

$$\begin{aligned}\ell_\Omega(\Omega) &= \sum_{i=1}^{n}\left\{\frac{\sum_{g_j\in\mathcal{G}_i} S_\Omega(D_i,X_i,g_j,\Omega)}{\sum_{g_j\in\mathcal{G}_i} S(D_i,X_i,g_j,\Omega)} - \frac{\sum_{d,j} S_\Omega(d,X_i,g_j,\Omega)}{\sum_{d,j} S(d,X_i,g_j,\Omega)}\right\}\\ &= \sum_{i=1}^{n}\{A_1(\Delta_i,D_i,X_i,G_i^*,\Omega) - A_2(X_i,\Omega)\}.\end{aligned}$$

Define $A_3(d,\Omega) = \mathrm{E}\{A_1(\Delta,D,X,G^*,\Omega) - A_2(X,\Omega)|D=d\}$. Then, $\sum_{i=1}^n A_3(D_i,\Omega) = 0$, because the score is unbiased. Thus we can write,

$$\ell_\Omega(\Omega) = \sum_{i=1}^{n}\left\{A_1(\Delta_i,D_i,X_i,G_i^*,\Omega) - A_2(X_i,\Omega) - A_3(D_i,\Omega)\right\}.$$

Notice that each of the terms in this sum is independent and zero mean. Then,

$$\begin{aligned}\mathrm{E}\{\ell_\Omega(\Omega)\ell_{\Omega^T}(\Omega)\} &= \sum_{i=1}^{n}\mathrm{E}\left[\{A_1(\Delta_i,D_i,X_i,G_i^*,\Omega) - A_2(X_i,\Omega) - A_3(D_i,\Omega)\}\right.\\ &\qquad\left.\times\{A_1(\Delta_i,D_i,X_i,G_i^*,\Omega) - A_2(X_i,\Omega) - A_3(D_i,\Omega)\}^T\right]\\ &= \sum_{i=1}^{n}\mathrm{E}\left[\{A_1(\Delta_i,D_i,X_i,G_i^*,\Omega) - A_2(X_i,\Omega)\}\right.\\ &\qquad\left.\times\{A_1(\Delta_i,D_i,X_i,G_i^*,\Omega) - A_2(X_i,\Omega)\}^T\right] - \sum_{i=1}^{n}A_3(D_i,\Omega)A_3(D_i,\Omega)^T.\end{aligned}$$

The first term can be written as $D_1 - D_2 - D_2^T + D_3$, where

$$\begin{aligned}D_1 &= \mathrm{E}\left[\sum_{i=1}^{n}\frac{\sum_{g_j\in\mathcal{G}_i} S_\Omega(D_i,X_i,g_j,\Omega)\{\sum_{g_j\in\mathcal{G}_i} S_\Omega(D_i,X_i,g_j,\Omega)\}^T}{\{\sum_{g_j\in\mathcal{G}_i} S(D_i,X_i,g_j,\Omega)\}^2}\right]\\ &= \frac{n_0}{\mathrm{pr}(D=0)}\int_x f_X(x)\sum_{d,j,\ell}\frac{\sum_{g_j\in\mathcal{G}_\ell} S_\Omega(d,x,g_j,\Omega)\{\sum_{g_j\in\mathcal{G}_\ell} S_\Omega(d,x,g_j,\Omega)\}^T}{\sum_{g_j\in\mathcal{G}_\ell} S(d,x,g_j,\Omega)}\\ &\quad\times\pi(\Delta|d,x,g_j^*)dx\\ D_2 &= \mathrm{E}\left[\sum_{i=1}^{n}\frac{\sum_{g_j\in\mathcal{G}_i} S_\Omega(D_i,X_i,g_j,\Omega)\{\sum_{d,j} S_\Omega(d,X_i,g_j,\Omega)\}^T}{\sum_{g_j\in\mathcal{G}_i} S(D_i,X_i,g_j,\Omega)\{\sum_{dj} S(d,X_i,g_j,\Omega)\}}\right]\\ &= \frac{n_0}{\mathrm{pr}(D=0)}\int_x f_X(x)\frac{\sum_{d,j} S_\Omega(d,x,g_j,\Omega)\{\sum_{d,j} S_\Omega(d,x,g_j,\Omega)\}^T}{\sum_{d,j} S(d,x,g_j,\Omega)}dx\\ D_3 &= \mathrm{E}\left[\sum_{i=1}^{n}\frac{\sum_{d,j} S_\Omega(d,X_i,g_j,\Omega)\{\sum_{d,j} S_\Omega(d,X_i,g_j,\Omega)\}^T}{\{\sum_{d,j} S(d,X_i,g_j,\Omega)\}^2}\right]\\ &= \frac{n_0}{P(D=0)}\int_x f_X(x)\frac{\sum_{d,j} S_\Omega(d,x,g_j,\Omega)\{\sum_{d,j} S_\Omega(d,x,g_j,\Omega)\}^T}{\sum_{d,j} S(d,x,g_j,\Omega)}dx,\end{aligned}$$

by repeated application of Lemma A.2. Notice that $D_2 = D_3$. Thus, the first term is just equal to $D_1 - D_3$. Additionally, $D_1 - D_3 = -\mathrm{E}\{\ell_{\Omega\Omega^T}(\Omega)]\}$. Then,

$$\mathrm{E}\{\ell_\Omega(\Omega)\ell_{\Omega^T}(\Omega)\} \;=\; -\mathrm{E}\{\ell_{\Omega\Omega^T}(\Omega)\} - \sum_{i=1}^{n} A_3(D_i,\Omega)A_3(D_i,\Omega)^T = n(I-\Lambda).$$

Let $R_d = (1/n_d)\sum_{d_i=d}\Psi(\Delta_i, D_i, X_i, G_i^*, \Omega)$. All of the elements in this sum are independent and identically distributed. Then, when n_0/n is constant, the central limit theorem implies that

$$n^{1/2}(n_d/n)^{1/2}\{R_d - \mathrm{E}(R_d)\} \Rightarrow \mathrm{N}\left[0, \mathrm{Var}\{\Psi(\Delta, D, X, G^*, \Omega)|D=d\}\right].$$

Now, since n_0/n is a constant, this implies that

$$n^{1/2}\{R_d - \mathrm{E}(R_d)\} \Rightarrow \mathrm{N}\left(0, n/n_d\mathrm{Var}\{\Psi(\Delta, D, X, G^*, \Omega)|D=d\}\right).$$

Also, the cases and the controls are independent and $(1/n)\ell_\Omega(\Omega) = (n_0/n)R_0 + (n_1/n)R_1$, so

$$n^{-1/2}\left[\ell_\Omega(\Omega) - \mathrm{E}\left\{\ell_\Omega(\Omega)\right\}\right] \Rightarrow \mathrm{N}\left[0, \sum_d n_d/n\mathrm{Var}\{\Psi(\Delta, D, X, G^*, \Omega)|D=d\}\right].$$

Next, $\mathrm{E}\{(1/n)\ell_\Omega(\Omega)\} = \{(1/n)\ell_\Omega(\Omega)\}|_{\Omega=\hat{\Omega}}$, and by the delta method

$$n^{1/2}(\hat{\Omega} - \Omega) \Rightarrow \mathrm{N}\left(0, I^{-1} - I^{-1}\Lambda I^{-1}\right).$$

APPENDIX C

PROOF FOR CHAPTER IV

C.1 The Profile Likelihood for Frequency Matched Data

For a fixed value of $\gamma = (\beta_0, \beta_1, \theta)$ the likelihood function for ζ has the form

$$\begin{aligned}
\ell(\zeta|\gamma) &\propto \sum_{i=1}^{n} \left[\log\{\zeta(X_i, S_i, W_i)\} - \log \left\{ \sum_{j,k,n} \text{pr}(D_i|x_k, s_n, w_m, g_j)\text{pr}(g_j|s_n)\zeta_{k,m,n} \right\} \right] \\
&= \sum_{k,m,n} [n_{kmn} \log\{\zeta(x_k, s_n, w_m)\}] \\
&\quad - \sum_{dm} \left[n_{dm} \log \left\{ \sum_{j,k,n} \text{pr}(D_i|x_k, s_n, w_m, g_j)\text{pr}(g_j|s_n)\zeta_{k,m,n} \right\} \right],
\end{aligned}$$

where $n_{dkmn} = \sum_{i=1}^{n} I(D_i = d, X_i = x_k, S_i = s_n, W_i = w_m)$. Then, taking derivatives,

$$\ell_{\zeta_{kmn}}(\zeta|\gamma) = \frac{n_{kmn}}{\zeta_{kmn}} - \sum_{d} \frac{n_{dm} \sum_j \{\text{pr}(d|g_j, x_k, s_n, w_m)\text{pr}(g_j|s_n)\}}{\sum_{j,k,n} \{\text{pr}(d|x_k, s_n, w_m, g_j)\text{pr}(g_j|s_n)\zeta_{k,m,n}\}}.$$

Setting the derivative equal to zero and solving for the MLE of the ζs gives

$$\begin{aligned}
\zeta_{kmn} &= n_{kmn} / \sum_{d} \frac{n_{dm} \sum_j \{\text{pr}(d|g_j, x_k, s_n, w_m)\text{pr}(g_j|s_n)\}}{\sum_{j,k,n} \{\text{pr}(d|x_k, s_n, w_m, g_j)\text{pr}(g_j|s_n)\zeta_{k,m,n}\}} \\
&= n_{kmn} / \sum_{d} \frac{n_{dm}\text{pr}(d|x_k, s_n, w_m)}{\text{pr}(d, w_m)}.
\end{aligned}$$

Define $\mu_{dm} = n_{dm}/\{n\text{pr}(D = d|W = w_m)\}$, which implies that $\text{pr}(D = d, W = w_m) = \{n_{dm}\text{pr}(W = w_m)\}/(n\mu_{dm})$. Then,

$$\begin{aligned}
\zeta_{kmn} &= n_{kmn} / \left\{ \sum_{d} \frac{n\mu_{dm}\text{pr}(d|x_k, s_n, w_m)}{\text{pr}(W = w_m)} \right\} \\
&= \frac{n_{kmn}\text{pr}(W = w_m)}{\sum_d n\mu_{dm}\text{pr}(d|x_k, s_n, w_m)}.
\end{aligned}$$

Then, the profile likelihood is

$$\begin{aligned}
\ell(\zeta(\gamma),\gamma) &= \sum_{i=1}^{n}\left[\log\left\{\sum_{g_j\in\mathcal{G}_i}\mathrm{pr}(D_i|g_j,X_i,S_i,W_i)h(g_j|S_i)\right\}\right] \\
&\quad + \sum_{k,m,n}\left\{n_{kmn}\log(\zeta_{kmn})\right\} - \sum_{dm}\left[n_{dm}\log\left\{\mathrm{pr}(d,w_m)\right\}\right] \\
&= \sum_{i=1}^{n}\left[\log\left\{\sum_{g_j\in\mathcal{G}_i}\mathrm{pr}(D_i|g_j,X_i,S_i,W_i)h(g_j|S_i)\right\}\right] \\
&\quad + \sum_{k,m,n}\left(n_{kmn}\left[\log\{\mathrm{pr}(W=w_m)\} - \log\{\sum_d \mu_{dm}\mathrm{pr}(d|x_k,s_n,w_m)\}\right]\right) \\
&\quad - \sum_{dm}\left(n_{dm}\left[\log\{\mathrm{pr}(W=w_m)\} + \log\{\mathrm{pr}(D=d|W=w_m)\}\right]\right) \\
&= \sum_{i=1}^{n}\left[\log\left\{\sum_{g_j\in\mathcal{G}_i}\mathrm{pr}(D_i|g_j,X_i,S_i,W_i)h(g_j|S_i)\right\}\right] \\
&\quad - \sum_{k,m,n}\left[n_{kmn}\log\{\sum_d \mu_{dm}\mathrm{pr}(d|x_k,s_n,w_m)\}\right] + \sum_{dm}\left\{n_{dm}\log(\mu_{dm})\right\}.
\end{aligned}$$

Now, define $\kappa_m = \beta_{0,m} + \log(\mu_{1m}/\mu_{0m})$ and $\Omega = (\gamma,\kappa)$. Then,

$$\begin{aligned}
\ell(\Omega) &= \sum_{i=1}^{n}\left[\log\left\{\sum_{g_j\in\mathcal{G}_i}\mathrm{pr}(D_i|g_j,X_i,S_i,W_i)h(g_j|S_i)\right\}\right] + \sum_{dm}\left[n_{dm}d(\kappa_m-\beta_{0m})\right] \\
&\quad - \sum_{k,m,n}\left[n_{kmn}\log\{\sum_d \exp\{d(\kappa_m-\beta_{0m})\}\mathrm{pr}(d|x_k,s_n,w_m)\}\right].
\end{aligned}$$

Let $S(D,G,X,S,W,\Omega) = \mathrm{pr}(D|G,X,S,W)\mathrm{pr}(G|S)\exp\{D(\kappa_W-\beta_{0W})\}$. Then, the profile likelihood has the form

$$\begin{aligned}
\ell(\Omega) &= \sum_{i=1}^{n}\left[\log\left\{\sum_{g_j\in\mathcal{G}_i}S(D_i,g_j,X_i,S_i,W_i,\Omega)\right\}\right. \\
&\qquad \left. - \log\left\{\sum_{d,j}S(d,g_j,X_i,S_i,W_i,\Omega)\right\}\right].
\end{aligned}$$

C.2 Modifications to Permit Analysis

For case (1), observe that when $\text{pr}(D = 1|W)$ is known for all W, the likelihood can be written in terms of $\Omega = (\beta_0, \beta_1, \theta)$ and fit just like a stratified population with a separate intercept term for each value of the matching variable.

APPENDIX D

PROOFS FOR CHAPTER V

D.1 The Pseudo-likelihood Function

The log of the pseudo-likelihood for and individual has the form

$$\begin{aligned}\ell(\Omega) &= \sum_{k=1}^{K}\Bigg[I(\Delta=k)\Bigg\{\log\{\pi(k|D,X,\xi)\}-\log\{\sum_{d,j}S(d,X,g_j,\Omega)\} \\ &\quad +\log\left(\sum_{\ell=1}^{L_k}\left[\sum_{g\in\zeta_{k\ell}}S(D,X,g,\Omega)\prod_{s=1}^{L_k}\{\pi_{\mathcal{M}k}(s|\zeta_{kl},D,\eta)\}^{I(\mathcal{M}=\zeta_{ks})}\right]\right)\Bigg\}\Bigg] \\ &\propto \sum_{k=1}^{K}I(\Delta=k)\log\left(\sum_{\ell=1}^{L_k}\left[\sum_{g\in\zeta_{k\ell}}S(D,X,g,\Omega)\prod_{s=1}^{L_k}\{\pi_{\mathcal{M}k}(s|\zeta_{kl},D,\eta)\}^{I(\mathcal{M}=\zeta_{ks})}\right]\right) \\ &\quad -\log\{\sum_{d,j}S(d,X,g_j,\Omega)\}\end{aligned}$$

Thus, by summing over all individuals, the complete likelihood has the claimed form.

D.2 The Asymptotic Distribution of the Estimates

D.2.1 The Score

Define $S_\Omega(D,X,G,\Omega)=\partial/\partial\Omega\{S(D,X,G,\Omega)\}$ and $\ell_\Omega(\Omega)=\partial/\partial\Omega\{\ell(\Omega)\}$. Now, the score of the profile likelihood has the form

$$\begin{aligned}\ell_\Omega(\Omega) &= \sum_{i,k}I(\Delta_i=k)\frac{\sum_{\ell,g\in\zeta_{k\ell}}S_\Omega(D_i,X_i,g,\Omega)\prod_{s=1}^{L_k}\{\pi_{\mathcal{M}_ik}(s|\zeta_{kl},D_i,\eta)\}^{I(\mathcal{M}_i=\zeta_{ks})}}{\sum_{\ell,g\in\zeta_{k\ell}}S(D_i,X_i,g,\Omega)\prod_{s=1}^{L_k}\{\pi_{\mathcal{M}_ik}(s|\zeta_{kl},D_i,\eta)\}^{I(\mathcal{M}_i=\zeta_{ks})}} \\ &\quad -\sum_{i=1}^{n}\frac{\sum_{d,j}S_\Omega(d,X_i,g_j,\Omega)}{\sum_{d,j}S(d,X_i,g_j,\Omega)}.\end{aligned}$$

Lemma A.1: The score of the profile likelihood is unbiased, and thus can be considered as a set of unbiased estimating equations.

The following lemma is useful in studying the distributional properties of the estimates obtained from the profile likelihood.

Lemma A.2: For any function $R(\Delta, D, X, \mathcal{M})$,

$$\mathrm{E}\left\{\sum_{i=1}^{n} R(\Delta_i, D_i, X_i, \mathcal{M}_i)\right\} = \frac{n_0}{\mathrm{pr}(D=0)} \int_x \sum_{d,k,l,r} \sum_{g \in \zeta_{kr}} R(\Delta = k, d, x, \mathcal{M} = \zeta_{kl}) \times f_X(x) S(d, x, g, \Omega) \pi(\Delta = k | d, x) \pi_{\mathcal{M},k}(\ell | \mathcal{G} = \zeta_{kr}, d, \eta) dx.$$

Proof of Lemma A.2: Notice,

$$\begin{aligned}\mathrm{E}\left\{\sum_{i=1}^{n} R(\Delta_i, D_i, X_i, \mathcal{M}_i)\right\} &= n_1 \int_x \sum_{k,\ell} R(\Delta = k, D = 1, x, \mathcal{M} = \zeta_{k\ell}) \\ &\quad \times f_{\Delta,X,\mathcal{M}|D}(\Delta = k, x, \zeta_{k\ell} | D = 1) dx \\ &\quad + n_0 \int_x \sum_{k,\ell} R(\Delta = k, D = 0, x, \mathcal{M} = \zeta_{k\ell}) f_{\Delta,X,\mathcal{M}|D}(\Delta = k, x, \zeta_{k\ell} | D = 0) dx.\end{aligned}$$

Also,

$$\begin{aligned} f_{\Delta,X,\mathcal{M}|D}(k, x, \zeta_{k\ell} | D) &= \sum_{r=1}^{L_k} f(X, \mathcal{M} = \zeta_{k\ell}, \Delta = k, \mathcal{G} = \zeta_{kr} | D) \\ &= \sum_{r} \pi_{\mathcal{M}k}(\ell | \mathcal{G}_k = \zeta_{kr}, D, \eta) \pi(k | D, X, \xi) f(\mathcal{G}_k, X | D). \end{aligned}$$

Additionally,

$$\begin{aligned} f(G = g, X = x | D = d) &= \frac{1}{\Pr(D=d)} \Pr(D = d, G = g, X = x) \\ &= \frac{1}{\Pr(D=d)} \Pr(D = d | G = g, X = x) \Pr(G = g) \\ &\quad \times \Pr(X = x) \\ &= \frac{1}{\Pr(D=0)} \left\{\frac{n_0}{n_1} \exp(\kappa - \beta_0)\right\}^d \Pr(G = g) \Pr(X = x) \\ &\quad \times \exp[d\{\beta_0 + m(g, x, \beta_1)\}] \left[1 - H\{\beta_0 + m(g, x, \beta_1)\}\right] \\ &= \frac{\Pr(X = x)}{\Pr(D=0)} \left(\frac{n_0}{n_1}\right)^d S(d, g, x, \Omega). \end{aligned}$$

Thus,

$$f(\mathcal{G}_k = \zeta_{kr}, X | D = d) = \frac{(n_0/n_1)^d}{\Pr(D=0)} f_X(x) \sum_{g \in \zeta_{kr}} S(d, x, g, \Omega).$$

Combining the above forms, the result immediately follows.

Proof of Lemma A.1: First, consider the second term in the score, which has form

$$A_2(X,\Omega) \;=\; \frac{\sum_{d,j} S_\Omega(d,X,g_j,\Omega)}{\sum_{d,j} S(d,X,g_j,\Omega)}.$$

¿From lemma A.2, this has

$$\begin{aligned}
\mathrm{E}\{\sum_i A_2(X_i,\Omega)\} &= \frac{n_0}{\mathrm{pr}(D=0)}\int_x f_X(x)\sum_{d,k,l,r}\sum_{g\in\zeta_{kr}}\frac{\sum_{d,j} S_\Omega(d,x,g_j,\Omega)}{\sum_{d,j} S(d,x,g_j,\Omega)} \\
&\quad \times S(d,x,g,\Omega)\pi(\Delta=k|d,x)\pi_{\mathcal{M},k}(\ell|\mathcal{G}=\zeta_{kr},d,\eta)dx \\
&= \frac{n_0}{\mathrm{pr}(D=0)}\int_x f_X(x)\sum_{d,k,r}\sum_{g\in\zeta_{kr}}\frac{\sum_{d,j} S_\Omega(d,x,g_j,\Omega)}{\sum_{d,j} S(d,x,g_j,\Omega)} \\
&\quad \times S(d,x,g,\Omega)\pi(\Delta=k|d,x)dx \\
&= \frac{n_0}{\mathrm{pr}(D=0)}\int_x f_X(x)\sum_{d,j,k}\frac{\sum_{d,j} S_\Omega(d,x,g_j,\Omega)}{\sum_{d,j} S(d,x,g_j,\Omega)} \\
&\quad \times S(d,x,g_j,\Omega)\pi(\Delta=k|d,x)dx \\
&= \frac{n_0}{\mathrm{pr}(D=0)}\int_x f_X(x)\sum_{d,j}\frac{\sum_{d,j} S_\Omega(d,x,g_j,\Omega)}{\sum_{d,j} S(d,x,g_j,\Omega)}S(d,x,g_j,\Omega)dx \\
&= \frac{n_0}{\mathrm{pr}(D=0)}\int_x f_X(x)\sum_{d,j} S_\Omega(d,x,g_j,\Omega)dx.
\end{aligned}$$

The first term is more difficult; it has the form

$$A_1(\Delta,D,X,\mathcal{M}) = \sum_{k=1}^{K} I(\Delta=k)\frac{\sum_{\ell,g} S_\Omega(D,X,g,\Omega)\prod_s \pi_{\mathcal{M}k}(s|\zeta_{kl},D,\eta)^{I(\mathcal{M}=\zeta_{ks})}}{\sum_{\ell,g} S(D,X,g,\Omega)\prod_s \pi_{\mathcal{M}k}(s|\zeta_{kl},D,\eta)^{I(\mathcal{M}=\zeta_{ks})}}$$

Then,

$$A_1(\Delta=k,D=d,X=x,\mathcal{M}=\zeta_{k\ell}) \;=\; \frac{\sum_{r=1}^{L_k}\sum_{g\in\zeta_{kr}} S_\Omega(d,x,g,\Omega)\pi_{\mathcal{M}k}(\ell|\zeta_{kr},d,\eta)}{\sum_{r=1}^{L_k}\sum_{g\in\zeta_{kr}} S(d,x,g,\Omega)\pi_{\mathcal{M}k}(\ell|\zeta_{kr},d,\eta)}$$

and

$$\begin{aligned}
\mathrm{E}\left\{\sum_i A_1(\cdot)\right\} &= \frac{n_0}{\mathrm{pr}(D=0)}\int_x \sum_{d,k,l,r}\sum_{g\in\zeta_{kr}} \frac{\sum_{r,g\in\zeta_{kr}} S_\Omega(d,x,g,\Omega)\pi_{\mathcal{M}k}(\ell|\zeta_{kr},d,\eta)}{\sum_{r,g\in\zeta_{kr}} S(d,x,g,\Omega)\pi_{\mathcal{M}k}(\ell|\zeta_{kr},d,\eta)} \\
&\quad \times f_X(x)S(d,x,g,\Omega)\pi(\Delta=k|d,x)\pi_{\mathcal{M},k}(\ell|\mathcal{G}=\zeta_{kr},d,\eta)dx \\
&= \frac{n_0}{\mathrm{pr}(D=0)}\int_x f_X(x)\sum_{d,k,l}\frac{\sum_{r,g\in\zeta_{kr}} S_\Omega(d,x,g,\Omega)\pi_{\mathcal{M}k}(\ell|\zeta_{kr},d,\eta)}{\sum_{r,g\in\zeta_{kr}} S(d,x,g,\Omega)\pi_{\mathcal{M}k}(\ell|\zeta_{kr},d,\eta)} \\
&\quad \times\pi(\Delta=k|d,x)\sum_r\sum_{g\in\zeta_{kr}} S(d,x,g,\Omega)\pi_{\mathcal{M},k}(\ell|\mathcal{G}=\zeta_{kr},d,\eta)dx \\
&= \frac{n_0}{\mathrm{pr}(D=0)}\int_x f_X(x)\sum_{d,k,l}\sum_{r=1}^{L_k}\sum_{g\in\zeta_{kr}} S_\Omega(d,x,g,\Omega)\pi_{\mathcal{M}k}(\ell|\zeta_{kr},d,\eta) \\
&\quad \times\pi(\Delta=k|d,x)dx \\
&= \frac{n_0}{\mathrm{pr}(D=0)}\int_x f_X(x)\sum_{d,k,r}\sum_{g\in\zeta_{kr}} S_\Omega(d,x,g,\Omega)\pi(\Delta=k|d,x)dx \\
&= \frac{n_0}{\mathrm{pr}(D=0)}\int_x f_X(x)\sum_{d,j,k} S_\Omega(d,x,g_j,\Omega)\pi(\Delta=k|d,x)dx \\
&= \frac{n_0}{\mathrm{pr}(D=0)}\int_x f_X(x)\sum_{d,j} S_\Omega(d,x,g_j,\Omega)dx
\end{aligned}$$

Thus, as both terms have the same expectation and opposite signs, $\mathrm{E}\{\ell_\Omega(\Omega)\}=0$.

In the case where the ηs are also unknown, the score with respect to η has the form

$$\ell_\eta(\Omega,\eta)=\sum_{k=1}^{K} I(\Delta=k)\frac{\sum_{\ell,g\in\zeta_{k\ell}} S(D,X,g,\Omega)\prod_{s=1}^{L_k}\{\pi_{\mathcal{M}k,\eta}(s|\zeta_{kl},D,\eta)\}^{I(\mathcal{M}=\zeta_{ks})}}{\sum_{\ell,g\in\zeta_{k\ell}} S(D,X,g,\Omega)\prod_{s=1}^{L_k}\{\pi_{\mathcal{M}k}(s|\zeta_{kl},D,\eta)\}^{I(\mathcal{M}=\zeta_{ks})}}.$$

This term is also unbiased, as

$$\begin{aligned}
\mathrm{E}\left\{\ell_\eta(\Omega,\eta)\right\} &= \frac{n_0}{\mathrm{pr}(D=0)}\int_x \sum_{d,k,l,r}\sum_{g\in\zeta_{kr}} \frac{\sum_{s=1}^{L_k}\sum_{g\in\zeta_{ks}} S(D,X,g,\Omega)\pi_{\mathcal{M}k,\eta}(\ell|\zeta_{ks},D,\eta)}{\sum_{s=1}^{L_k}\sum_{g\in\zeta_{ks}} S(D,X,g,\Omega)\pi_{\mathcal{M}k}(\ell|\zeta_{ks},D,\eta)} \\
&\quad \times f_X(x)S(d,x,g,\Omega)\pi(\Delta=k|d,x)\pi_{\mathcal{M},k}(\ell|\mathcal{G}=\zeta_{kr},d,\eta)dx \\
&= \frac{n_0}{\mathrm{pr}(D=0)}\int_x f_X(x)\sum_{d,k,l}\sum_{s=1}^{L_k}\sum_{g\in\zeta_{ks}} S(D,X,g,\Omega)\pi_{\mathcal{M}k,\eta}(\ell|\zeta_{ks},D,\eta) \\
&\quad \times\pi(\Delta=k|d,x)dx \\
&= 0,
\end{aligned}$$

as $\sum_\ell \pi_{\mathcal{M},k}(\ell|\mathcal{G} = \zeta_{kr}, d, \eta) = 0$. Similar arguments show that the variance of the score also has the desired form.

D.2.2 The Variance of the Score

Lemma A.3: For the case–control study design, with fixed n_0/n, measurable functions $R(\Delta, D, X, \mathcal{M})$ satisfy

$$\frac{1}{n}\sum_{i=1}^{n} R(\Delta_i, D_i, X_i, \mathcal{M}_i^*) \longrightarrow_P \mu_0 \int_x f_X(x) \sum_{d,k,\ell,r} \sum_{g\in\zeta_{kr}} R(\Delta = k, d, x, \mathcal{M} = \zeta_{k\ell}) \times S(d, x, g, \Omega)\pi(\Delta = k|d, x)\pi_{\mathcal{M},k}(\ell|\mathcal{G} = \zeta_{kr}, d, \eta)dx.$$

assuming that the integral exists.

Proof of Lemma 3: Notice

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} R(\Delta_i, D_i, X_i, \mathcal{M}_i) &= \frac{1}{n_0}\frac{n_0}{n}\sum_{i:d_i=0} R(\Delta_i, D_i = 0, X_i, \mathcal{M}_i) \\ &\quad + \frac{1}{n_1}\frac{n_1}{n}\sum_{i:d_i=1} R(\Delta_i, D_i = 1, X_i, \mathcal{M}_i).\end{aligned}$$

Now, the cases, $d_i = 1$, are iid from the distribution of cases, and the controls, $d_i = 0$, are iid from their distribution. Also,

$$\frac{1}{n_d}\sum_{i:d_i=d} R(\Delta_i, D_i, X_i, \mathcal{M}_i) \longrightarrow_P \mathrm{E}\{R(\Delta, D, X, \mathcal{M})|D = d\}$$

by the Weak Law of Large Numbers. Thus

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} R(\Delta_i, D_i, X_i, \mathcal{M}_i) \longrightarrow_P &\frac{n_0}{n}\mathrm{E}\{R(\Delta, D, X, \mathcal{M})|D = 0\} \\ &+ \frac{n_1}{n}\mathrm{E}\{R(\Delta, D, X, \mathcal{M})|D = 1\}.\end{aligned}$$

Also, from Lemma A.1,

$$\begin{aligned}\mathrm{E}\{R(\Delta, D, X, \mathcal{M})|D = d\} &= \frac{n\mu_0}{n_d}\int_x f_X(x)\sum_{k,\ell,r}\sum_{g\in\zeta_{kr}} R(\Delta = k, d, x, \mathcal{M} = \zeta_{k\ell}) \\ &\quad \times S(d, x, g_j, \Omega)\pi(\Delta = k|d, x)\pi_{\mathcal{M},k}(\ell|\mathcal{G} = \zeta_{kr}, d, \eta)dx.\end{aligned}$$

Thus,

$$\frac{1}{n}\sum_{i=1}^{n} R(\Delta_i, D_i, X_i, \mathcal{M}_i) \quad \longrightarrow_P \quad \mu_0 \int_x f_X(x) \sum_{d,k,\ell,r} \sum_{g\in\zeta_{kr}} R(\Delta = k, d, x, \mathcal{M} = \zeta_{k\ell}) \times S(d, x, g_j, \Omega)\pi(\Delta = k|d, x)\pi_{\mathcal{M},k}(\ell|\mathcal{G} = \zeta_{kr}, d, \eta)dx.$$

The Matrix of Second Partials: Define $S_{\Omega\Omega^T}(D, X, G, \Omega)$ to be the matrix of second derivatives of $S(D, X, G, \Omega)$ with respect to Ω. Define $\ell_{\Omega\Omega^T}(\Omega)$ similarly. First note that

$$\begin{aligned}
\ell_{\Omega\Omega^T}(\Omega) = \sum_{i=1}^{n} \Bigg\{ & \frac{\sum_{d,j} S_\Omega(d, X_i, g_j, \Omega)\{\sum_{d,j} S_\Omega(d, X_i, g_j, \Omega)\}^T}{\{\sum_{dj} S(d, X_i, g_j, \Omega)\}^2} \\
& - \frac{\sum_{d,j} S_{\Omega\Omega^T}(d, X_i, g_j, \Omega)}{\sum_{dj} S(d, X_i, g_j, \Omega)} \\
& + \sum_{k=1}^{K} I(\Delta_i = k) \Bigg(\frac{\sum_{\ell,g} S_{\Omega\Omega^T}(D_i, X_i, g, \Omega) \prod_s \{\pi_{\mathcal{M}_i k}(s|\zeta_{kl}, D_i, \eta)\}^{I(\mathcal{M}_i=\zeta_{ks})}}{\sum_{\ell,g} S(D_i, X_i, g, \Omega) \prod_s \{\pi_{\mathcal{M}_i k}(s|\zeta_{kl}, D_i, \eta)\}^{I(\mathcal{M}_i=\zeta_{ks})}} \\
& - \frac{\sum_{\ell,g\in\zeta_{k\ell}} S_\Omega(D_i, X_i, g, \Omega) \prod_{s=1}^{L_k} \{\pi_{\mathcal{M}_i k}(s|\zeta_{kl}, D_i, \eta)\}^{I(\mathcal{M}_i=\zeta_{ks})}}{\left[\sum_{\ell=1}^{L_k} \sum_{g\in\zeta_{k\ell}} S(D_i, X_i, g, \Omega) \prod_{s=1}^{L_k} \{\pi_{\mathcal{M}_i k}(s|\zeta_{kl}, D_i, \eta)\}^{I(\mathcal{M}_i=\zeta_{ks})}\right]^2} \\
& \times \left[\sum_{\ell,g\in\zeta_{k\ell}} S_\Omega(D_i, X_i, g, \Omega) \prod_{s=1}^{L_k} \{\pi_{\mathcal{M}_i k}(s|\zeta_{kl}, D_i, \eta)\}^{I(\mathcal{M}_i=\zeta_{ks})}\right]^T \Bigg) \Bigg\}.
\end{aligned}$$

Consider the second and third terms denoted by S_{n2} and S_{n3}, respectively. Then,

$$\begin{aligned}
\mathrm{E}(S_{n2}) &= \frac{n_0}{\mathrm{pr}(D=0)} \int_x f_X(x) \sum_{d,k,l,r} \sum_{g\in\zeta_{kr}} \frac{\sum_{d,j} S_{\Omega\Omega^T}(d, X_i, g_j, \Omega)}{\sum_{dj} S(d, X_i, g_j, \Omega)} \\
&\quad \times S(d, x, g, \Omega)\pi(\Delta = k|d, x)\pi_{\mathcal{M},k}(\ell|\mathcal{G} = \zeta_{kr}, d, \eta)dx \\
&= \frac{n_0}{\mathrm{pr}(D=0)} \int_x f_X(x) \sum_{d,j,k} \frac{\sum_{d,j} S_{\Omega\Omega^T}(d, X_i, g_j, \Omega)}{\sum_{dj} S(d, X_i, g_j, \Omega)} \\
&\quad \times S(d, x, g_j, \Omega)\pi(\Delta = k|d, x)dx \\
&= \frac{n_0}{\mathrm{pr}(D=0)} \int_x f_X(x) \sum_{d,j} S_{\Omega\Omega^T} S(d, X_i, g_j, \Omega)dx
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}(S_{n3}) &= \frac{n_0}{\mathrm{pr}(D=0)}\int_x \sum_{d,k,l,r}\sum_{g\in\zeta_{kr}} \frac{\sum_{r=1}^{L_k}\sum_{g\in\zeta_{kr}} S_{\Omega\Omega^T}(d,x,g,\Omega)\pi_{\mathcal{M}k}(\ell|\zeta_{kr},d,\eta)}{\sum_{r=1}^{L_k}\sum_{g\in\zeta_{kr}} S(d,x,g,\Omega)\pi_{\mathcal{M}k}(\ell|\zeta_{kr},d,\eta)} \\
&\quad \times f_X(x)S(d,x,g,\Omega)\pi(\Delta=k|d,x)\pi_{\mathcal{M},k}(\ell|\mathcal{G}=\zeta_{kr},d,\eta)dx \\
&= \frac{n_0}{\mathrm{pr}(D=0)}\int_x f_X(x)\sum_{d,k,l} \frac{\sum_{r=1}^{L_k}\sum_{g\in\zeta_{kr}} S_{\Omega\Omega^T}(d,x,g,\Omega)\pi_{\mathcal{M}k}(\ell|\zeta_{kr},d,\eta)}{\sum_{r=1}^{L_k}\sum_{g\in\zeta_{kr}} S(d,x,g,\Omega)\pi_{\mathcal{M}k}(\ell|\zeta_{kr},d,\eta)} \\
&\quad \times\pi(\Delta=k|d,x)\sum_r\sum_{g\in\zeta_{kr}} S(d,x,g,\Omega)\pi_{\mathcal{M},k}(\ell|\mathcal{G}=\zeta_{kr},d,\eta)dx \\
&= \frac{n_0}{\mathrm{pr}(D=0)}\int_x f_X(x)\sum_{d,k,l}\sum_{r=1}^{L_k}\sum_{g\in\zeta_{kr}} S_{\Omega\Omega^T}(d,x,g,\Omega)\pi_{\mathcal{M}k}(\ell|\zeta_{kr},d,\eta) \\
&\quad \times\pi(\Delta=k|d,x)dx \\
&= \frac{n_0}{\mathrm{pr}(D=0)}\int_x f_X(x)\sum_{d,k,r}\sum_{g\in\zeta_{kr}} S_{\Omega\Omega^T}(d,x,g,\Omega)\pi(\Delta=k|d,x)dx \\
&= \frac{n_0}{\mathrm{pr}(D=0)}\int_x f_X(x)\sum_{d,j,k} S_{\Omega\Omega^T}S(d,x,g_j,\Omega)\pi(\Delta=k|d,x)dx \\
&= \frac{n_0}{\mathrm{pr}(D=0)}\int_x f_X(x)\sum_{d,j} S_{\Omega\Omega^T}(d,x,g_j,\Omega)dx.
\end{aligned}$$

Then $\mathrm{E}(S_{n1}-S_{n2})=0$, and

$$\begin{aligned}
-\mathrm{E}\{\ell_{\Omega\Omega^T}(\Omega)\} &= -\mathrm{E}\left[\sum_{i=1}^{n}\frac{\sum_{d,j}S_\Omega(d,X_i,g_j,\Omega)\{\sum_{d,j}S_\Omega(d,X_i,g_j,\Omega)\}^T}{\{\sum_{dj}S(d,X_i,g_j,\Omega)\}^2}\right] \\
&\quad +\mathrm{E}\left(\sum_{i,k} I(\Delta_i=k)\frac{\sum_{\ell,g\in\zeta_{k\ell}}S_\Omega(D_i,X_i,g,\Omega)\prod_{s=1}^{L_k}\{\pi_{\mathcal{M}_i k}(s|\zeta_{kl},D_i,\eta)\}^{I(\mathcal{M}_i=\zeta_{ks})}}{[\sum_{\ell,g\in\zeta_{k\ell}}S(D_i,X_i,g,\Omega)\prod_{s=1}^{L_k}\{\pi_{\mathcal{M}_i k}(s|\zeta_{kl},D_i,\eta)\}^{I(\mathcal{M}_i=\zeta_{ks})}]^2}\right. \\
&\quad \left.\times\left[\sum_{\ell,g\in\zeta_{k\ell}}S_\Omega(D_i,X_i,g,\Omega)\prod_{s=1}^{L_k}\{\pi_{\mathcal{M}_i k}(s|\zeta_{kl},D_i,\eta)\}^{I(\mathcal{M}_i=\zeta_{ks})}\right]^T\right) \\
&= -\mathrm{E}(C_{n1})+\mathrm{E}(C_{n2}),
\end{aligned}$$

where

$$\begin{aligned}
\mathrm{E}(C_{n1}) &= \frac{n_0}{\mathrm{pr}(D=0)}\int_x \sum_{d,k,l,r}\sum_{g\in\zeta_{kr}} \frac{\sum_{d,j} S_\Omega(d,X_i,g_j,\Omega)\{\sum_{d,j} S_\Omega(d,X_i,g_j,\Omega)\}^T}{\{\sum_{dj} S(d,X_i,g_j,\Omega)\}^2} \\
&\quad \times f_X(x) S(d,x,g,\Omega)\pi(\Delta=k|d,x)\pi_{\mathcal{M},k}(\ell|\mathcal{G}=\zeta_{kr},d,\eta)dx \\
&= \frac{n_0}{\mathrm{pr}(D=0)}\int_x f_X(x)\sum_{d,j,k} \frac{\sum_{d,j} S_\Omega(d,X_i,g_j,\Omega)\{\sum_{d,j} S_\Omega(d,X_i,g_j,\Omega)\}^T}{\{\sum_{dj} S(d,X_i,g_j,\Omega)\}^2} \\
&\quad \times S(d,x,g_j,\Omega)\pi(\Delta=k|d,x)dx \\
&= \frac{n_0}{\mathrm{pr}(D=0)}\int_x f_X(x)\frac{\sum_{d,j} S_\Omega(d,X_i,g_j,\Omega)\{\sum_{d,j} S_\Omega(d,X_i,g_j,\Omega)\}^T}{\sum_{dj} S(d,X_i,g_j,\Omega)}dx
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}(C_{n2}) &= \frac{n_0}{\mathrm{pr}(D=0)}\int_x f_X(x)\sum_{d,k,l,r}\sum_{g\in\zeta_{kr}} \frac{\sum_r\sum_{g\in\zeta_{kr}} S_\Omega(d,x,g,\Omega)\pi_{\mathcal{M}k}(\ell|\zeta_{kr},d,\eta)}{\{\sum_r\sum_{g\in\zeta_{kr}} S(d,x,g,\Omega)\pi_{\mathcal{M}k}(\ell|\zeta_{kr},d,\eta)\}^2} \\
&\quad \times\left\{\sum_r\sum_{g\in\zeta_{kr}} S_\Omega(d,x,g,\Omega)\pi_{\mathcal{M}k}(\ell|\zeta_{kr},d,\eta)\right\}^T \\
&\quad \times S(d,x,g,\Omega)\pi(\Delta=k|d,x)\pi_{\mathcal{M},k}(\ell|\mathcal{G}=\zeta_{kr},d,\eta)dx \\
&= \frac{n_0}{\mathrm{pr}(D=0)}\int_x f_X(x)\sum_{d,k,l} \frac{\sum_r\sum_{g\in\zeta_{kr}} S_\Omega(d,x,g,\Omega)\pi_{\mathcal{M}k}(\ell|\zeta_{kr},d,\eta)}{\left\{\sum_r\sum_{g\in\zeta_{kr}} S(d,x,g,\Omega)\pi_{\mathcal{M}k}(\ell|\zeta_{kr},d,\eta)\right\}^2} \\
&\quad \times\left\{\sum_r\sum_{g\in\zeta_{kr}} S_\Omega(d,x,g,\Omega)\pi_{\mathcal{M}k}(\ell|\zeta_{kr},d,\eta)\right\}^T \\
&\quad \times\pi(\Delta=k|d,x)\sum_r\sum_{g\in\zeta_{kr}} S(d,x,g,\Omega)\pi_{\mathcal{M},k}(\ell|\mathcal{G}=\zeta_{kr},d,\eta)dx \\
&= \frac{n_0}{\mathrm{pr}(D=0)}\int_x f_X(x)\sum_{d,k,l} \frac{\sum_r\sum_{g\in\zeta_{kr}} S_\Omega(d,x,g,\Omega)\pi_{\mathcal{M}k}(\ell|\zeta_{kr},d,\eta)}{\sum_r\sum_{g\in\zeta_{kr}} S(d,x,g,\Omega)\pi_{\mathcal{M}k}(\ell|\zeta_{kr},d,\eta)} \\
&\quad \times\left\{\sum_r\sum_{g\in\zeta_{kr}} S_\Omega(d,x,g,\Omega)\pi_{\mathcal{M}k}(\ell|\zeta_{kr},d,\eta)\right\}^T \pi(\Delta=k|d,x)dx.
\end{aligned}$$

The Variance of the Score: Recall,

$$\begin{aligned}\ell_\Omega(\Omega) &= \sum_{i,k} I(\Delta_i = k)\frac{\sum_{\ell,g\in\zeta_{k\ell}} S_\Omega(D_i,X_i,g,\Omega)\prod_{s=1}^{L_k}\{\pi_{\mathcal{M}_i k}(s|\zeta_{kl},D_i,\eta)\}^{I(\mathcal{M}_i=\zeta_{ks})}}{\sum_{\ell,g\in\zeta_{k\ell}} S(D_i,X_i,g,\Omega)\prod_{s=1}^{L_k}\{\pi_{\mathcal{M}_i k}(s|\zeta_{kl},D_i,\eta)\}^{I(\mathcal{M}_i=\zeta_{ks})}} \\ &\quad -\sum_{i=1}^{n}\frac{\sum_{d,j} S_\Omega(d,X_i,g_j,\Omega)}{\sum_{d,j} S(d,X_i,g_j,\Omega)} \\ &= \sum_{i=1}^{n}\{A_1(\Delta_i,D_i,X_i,\mathcal{M}_i,\Omega) - A_2(X_i,\Omega)\}.\end{aligned}$$

Define $A_3(d,\Omega) = \mathrm{E}[\{A_1(\Delta,D,X,\mathcal{M},\Omega) - A_2(X,\Omega)\}|D=d]$. Then, as the score is unbiased, $\sum_{i=1}^n A_3(D_i,\Omega) = 0$. Thus we can write,

$$\ell_\Omega(\Omega) = \sum_{i=1}^{n}\{A_1(\Delta_i,D_i,X_i,\mathcal{M}_i,\Omega) - A_2(X_i,\Omega) - A_3(D_i,\Omega)\}.$$

Notice that each of the terms in this sum is independent and zero mean. Then,

$$\begin{aligned}\mathrm{E}\{\ell_\Omega(\Omega)\ell_{\Omega^T}(\Omega)\} &= \sum_{i=1}^{n}\mathrm{E}\left[\{A_1(\Delta_i,D_i,X_i,\mathcal{M}_i,\Omega) - A_2(X_i,\Omega) - A_3(D_i,\Omega)\}\right. \\ &\qquad \left.\times\{A_1(\Delta_i,D_i,X_i,\mathcal{M}_i,\Omega) - A_2(X_i,\Omega) - A_3(D_i,\Omega)\}^T\right] \\ &= \sum_{i=1}^{n}\mathrm{E}\left[\{A_1(\Delta_i,D_i,X_i,\mathcal{M}_i,\Omega) - A_2(X_i,\Omega)\}\right. \\ &\qquad \left.\times\{A_1(\Delta_i,D_i,X_i,\mathcal{M}_i,\Omega) - A_2(X_i,\Omega)\}^T\right] - \sum_{i=1}^{n} A_3(D_i,\Omega)A_3(D_i,\Omega)^T.\end{aligned}$$

The first term can be written as $D_1 - D_2 - D_2^T + D_3$, where

$$\begin{aligned}D_1 &= \mathrm{E}\left(\sum_{i,k} I(\Delta_i = k)\frac{\sum_{\ell,g\in\zeta_{k\ell}} S_\Omega(D_i,X_i,g,\Omega)\prod_s\{\pi_{\mathcal{M}_i k}(s|\zeta_{kl},D_i,\eta)\}^{I(\mathcal{M}_i=\zeta_{ks})}}{[\sum_{\ell,g\in\zeta_{k\ell}} S(D_i,X_i,g,\Omega)\prod_s\{\pi_{\mathcal{M}_i k}(s|\zeta_{kl},D_i,\eta)\}^{I(\mathcal{M}_i=\zeta_{ks})}]^2}\right. \\ &\qquad \left.\times\left[\sum_{\ell,g\in\zeta_{k\ell}} S_\Omega(D_i,X_i,g,\Omega)\prod_{s=1}^{L_k}\{\pi_{\mathcal{M}_i k}(s|\zeta_{kl},D_i,\eta)\}^{I(\mathcal{M}_i=\zeta_{ks})}\right]^T\right) \\ &= \frac{n_0}{\mathrm{pr}(D=0)}\int_x f_X(x)\sum_{d,k,l}\frac{\sum_r\sum_{g\in\zeta_{kr}} S_\Omega(d,x,g,\Omega)\pi_{\mathcal{M}k}(\ell|\zeta_{kr},d,\eta)}{\sum_r\sum_{g\in\zeta_{kr}} S(d,x,g,\Omega)\pi_{\mathcal{M}k}(\ell|\zeta_{kr},d,\eta)} \\ &\qquad \times\left\{\sum_r\sum_{g\in\zeta_{kr}} S_\Omega(d,x,g,\Omega)\pi_{\mathcal{M}k}(\ell|\zeta_{kr},d,\eta)\right\}^T \pi(\Delta=k|d,x)dx,\end{aligned}$$

$$
\begin{aligned}
D_2 &= \mathrm{E}\left[\sum_{ik} I(\Delta_i = k)\frac{\sum_{\ell,g\in\zeta_{k\ell}} S_\Omega(D_i, X_i, g, \Omega)\prod_{s=1}^{L_k}\{\pi_{\mathcal{M}_i k}(s|\zeta_{kl}, D_i, \eta)\}^{I(\mathcal{M}_i=\zeta_{ks})}}{\sum_{\ell,g\in\zeta_{k\ell}} S(D_i, X_i, g, \Omega)\prod_{s=1}^{L_k}\{\pi_{\mathcal{M}_i k}(s|\zeta_{kl}, D_i, \eta)\}^{I(\mathcal{M}_i=\zeta_{ks})}}\right. \\
&\quad \left.\times \frac{\{\sum_{d,j} S_\Omega(d, X_i, g_j, \Omega)\}^T}{\{\sum_{dj} S(d, X_i, g_j, \Omega)\}}\right] \\
&= \frac{n_0}{\mathrm{pr}(D=0)}\int_x \sum_{d,k,l,r,g} \frac{\sum_{r,g} S_\Omega(d, x, g, \Omega)\pi_{\mathcal{M}k}(\ell|\zeta_{kr}, d, \eta)\{\sum_{d,j} S_\Omega(d, x, g_j, \Omega)\}^T}{\sum_{r,g} S(d, x, g, \Omega)\pi_{\mathcal{M}k}(\ell|\zeta_{kr}, d, \eta)\{\sum_{dj} S(d, X_i, g_j, \Omega)\}} \\
&\quad \times f_X(x) S(d, x, g, \Omega)\pi(\Delta = k|d, x)\pi_{\mathcal{M},k}(\ell|\mathcal{G} = \zeta_{kr}, d, \eta)dx \\
&= \frac{n_0}{\mathrm{pr}(D=0)}\int_x \sum_{d,k,l} \frac{\sum_{r,g} S_\Omega(d, x, g, \Omega)\pi_{\mathcal{M}k}(\ell|\zeta_{kr}, d, \eta)\{\sum_{d,j} S_\Omega(d, x, g_j, \Omega)\}^T}{\sum_{dj} S(d, X_i, g_j, \Omega)} \\
&\quad \times f_X(x)\pi(\Delta = k|d, x)dx \\
&= \frac{n_0}{\mathrm{pr}(D=0)}\int_x f_X(x) \sum_{d,k} \frac{\sum_{r,g\in\zeta_{kr}} S_\Omega(d, x, g, \Omega)\{\sum_{d,j} S_\Omega(d, x, g_j, \Omega)\}^T}{\sum_{dj} S(d, X_i, g_j, \Omega)} \\
&\quad \times \pi(\Delta = k|d, x)dx \\
&= \frac{n_0}{\mathrm{pr}(D=0)}\int_x f_X(x)\frac{\sum_{d,j} S_\Omega(d, x, g, \Omega)\{\sum_{d,j} S_\Omega(d, x, g_j, \Omega)\}^T}{\sum_{dj} S(d, X_i, g_j, \Omega)}dx,
\end{aligned}
$$

and

$$
\begin{aligned}
D_3 &= \mathrm{E}\left[\sum_{i=1}^{n} \frac{\sum_{d,j} S_\Omega(d, X_i, g_j, \Omega)\{\sum_{d,j} S_\Omega(d, X_i, g_j, \Omega)\}^T}{\{\sum_{d,j} S(d, X_i, g_j, \Omega)\}^2}\right] \\
&= \frac{n_0}{P(D=0)}\int_x f_X(x)\frac{\sum_{d,j} S_\Omega(d, x, g_j, \Omega)\{\sum_{d,j} S_\Omega(d, x, g_j, \Omega)\}^T}{\sum_{d,j} S(d, x, g_j, \Omega)}dx
\end{aligned}
$$

by repeated application of Lemma A.2. Notice that $D_2 = D_3$. Thus, the first term is just equal to $D_1 - D_3$. Additionally, $D_1 - D_3 = -\mathrm{E}\{\ell_{\Omega\Omega^T}(\Omega)\}$. Then,

$$
\begin{aligned}
\mathrm{E}\{\ell_\Omega(\Omega)\ell_{\Omega^T}(\Omega)\} &= -\mathrm{E}\{\ell_{\Omega\Omega^T}(\Omega)\} - \sum_{i=1}^{n} A_3(D_i, \Omega)A_3(D_i, \Omega)^T \\
&= I - \Lambda.
\end{aligned}
$$

Let $R_d = (1/n_d)\sum_{d_i=d}\Psi(\Delta_i, D_i, X_i, \mathcal{M}_i, \Omega)$. All of the elements in this sum are independent and identically distributed. Then, when n_0/n is constant, the central limit theorem implies that

$$
n^{1/2}(n_d/n)^{1/2}\{R_d - \mathrm{E}(R_d)\} \Rightarrow \mathrm{N}\left(0, \mathrm{Var}\{\Psi(\Delta, D, X, \mathcal{M}, \Omega)|D = d\}\right).
$$

Now, since n_0/n is a constant, this implies that

$$n^{1/2}\{R_d - \mathrm{E}(R_d)\} \Rightarrow \mathrm{N}\left(0, \frac{n}{n_d}\mathrm{Var}\{\Psi(\Delta, D, X, \mathcal{M}, \Omega)|D = d\}\right).$$

Also, the cases and the controls are independent and $(1/n)\ell_\Omega(\Omega) = (n_0/n)R_0 + (n_1/n)R_1$, so

$$n^{1/2}\left[\frac{1}{n}\ell_\Omega(\Omega) - \mathrm{E}\left\{\frac{1}{n}\ell_\Omega(\Omega)\right\}\right] \Rightarrow \mathrm{N}\left(0, \sum_d \frac{n_d}{n}\mathrm{Var}\{\Psi(\Delta, D, X, \mathcal{M}, \Omega)|D = d\}\right).$$

Next, $\mathrm{E}\{(1/n)\ell_\Omega(\Omega)\} = \{(1/n)\ell_\Omega(\Omega)\}|_{\Omega=\hat{\Omega}}$, and by the delta method

$$n^{1/2}(\hat{\Omega} - \Omega) \Rightarrow \mathrm{N}\left(0, I^{-1} - I^{-1}\Lambda I^{-1}\right).$$

VITA

Christine Spinka was born in Hinsdale, Illinois. She received a Bachelor of Science degree in mathematics and molecular biology in 1999 from Vanderbilt University. In December 2003, she received a Master of Science degree in statistics, and in December 2004 she received a Ph.D. degree in statistics, both from Texas A&M University. Her permanent address is:

3000 Trailside Dr.

Columbia, MO 65203